DOCUMENT RESUME

ED 151 414 95 TM 006 943

AUTHOR Subkoviak, Michael J.
TITLE Evaluation of Criterion-Referenced Reliability Coefficients. Final Report.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C. Basic Skills Group.
BUREAU NO 6-0357
PUB DATE 31 Dec 77
GRANT NIE-G-76-0088
NOTE 113p.

EDRS PRICE MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; *Cutting Scores; *Data Analysis; Data Collection; *Mastery Tests; *Mathematical Models; Research Design; Research Methodology; Sampling; Statistical Analysis; Statistical Data; *Test Reliability
IDENTIFIERS Test Length

ABSTRACT

Four different procedures were used for estimating the proportion of persons who would be classified consistently as either passing both of two parallel tests or failing both. These four methods were applied at each of four different mastery level scores for each of three different length tests. Data were based on 50 replications of each procedure for samples of 30 cases and 300 cases randomly drawn from a population of 1586 cases. The outcomes of these sampling experiments were compared with the parameter values for the population. In addition to this study, four other papers on related topics are attached. "Empirical Investigation of Procedures for Estimating Reliability for Mastery Tests" discusses the data presented in the previous paper. The next paper, "Estimating the Probability of Correct Classification in Mastery Testing," discusses the Keats-Lord model used in the preceding papers. "Further Comments on Reliability for Mastery Tests" explores the use of coefficients of classification consistency. The last paper, "Confirmatory Inference and Geometric Models," discusses the relationship between exploratory and confirmatory approaches to collecting and analyzing data and the application of geometric models to these approaches. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED151414

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

FINAL REPORT

PROJECT NO. 6-0357
GRANT NO. NIE-G-76-0088

EVALUATION OF CRITERION-REFERENCED RELIABILITY COEFFICIENTS

MICHAEL J. SUBKOVIAK
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN 53706

December 31, 1977

The research reported herein was performed pursuant to a grant with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to freely express their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official National Institute of Education position or policy.

U.S. DEPARTMENT OF

HEALTH, EDUCATION AND WELFARE

NATIONAL INSTITUTE OF EDUCATION

BASIC SKILLS

TM006 943

Abstract

Four different procedures (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976; Swaminathan, Hambleton & Algina, 1974) have been proposed for estimating the proportion of persons consistently classified as master/master or nonmaster/nonmaster on two mastery tests. Estimates of this proportion were obtained by each of the above procedures for repeated samples of 30 and 300 persons drawn from a population of 1586. These estimates were then compared for accuracy to the proportion of the population of 1586 consistently classified as master/master or nonmaster/nonmaster on two tests; hereafter, this proportion is referred to as the population parameter. Reasonably accurate estimates of the population parameter were generally obtained for all four procedures; however, instances of systematic estimation bias were observed, especially for tests of 10 items or less. For example, the Huynh procedure tended to produce underestimates of the population parameter, while the Marshall-Haertel and Subkoviak procedures produced underestimates in certain instances and overestimates in others. The Swaminathan-Hambleton-Algina procedure generally produced unbiased estimates, however these estimates tended to deviate widely from the population parameter, especially for samples of 30 persons or less.

EVALUATION OF CRITERION-REFERENCED RELIABILITY COEFFICIENTS

MICHAEL J. SUBKOVIAK
UNIVERSITY OF WISCONSIN

Introduction

For present purposes, a mastery test can be defined as a test with a single cutting score, c, that determines mastery and nonmastery subgroups, having scores above and below c respectively. In this context, reliability refers to the consistency of mastery-nonmastery decisions over repeated test administrations (Hambleton & Novick, 1973, pp. 166-167). Accordingly, the proportion of consistent mastery/mastery and nonmastery/nonmastery classifications for a group on two tests with cutting score c, sympolized P_c, has been proposed as a raw index of reliability in this context (Swaminathan, Hambleton & Algina, 1974, 1975). In addition, three procedures for estimating proportion P_c from scores on a single test have emerged (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976). Thus, a teacher or other test user is faced with the problem of choosing among four different procedures for estimating P_c in the absence of any clear guidelines. The purpose of the project was to provide such guidelines by comparing population values of P_c to sample estimates of P_c. Specifically, the proportion of consistent classifications on two tests for a population of 1586, a P_c - parameter, was compared for accuracy to four different P_c - estimates (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976; Swaminathan et al., 1974) for repeated samples of 30 and 300 persons from the same population. The results thus illustrate the extent and nature of discrepancies between parameter and estimates. The results also have indirect relevance for the process of estimating coefficient kappa, a function of P_c that has also been proposed as an index of reliability for mastery tests (see Huynh, 1976; Subkoviak, 1977; Swaminathan et al., 1974).

Method

The data base for the project consisted of the responses of a population of 1586 students to parallel tests of 10, 30, and 50 items each from the Scholastic Aptitude Test. Each 10-item test was part of the longer 30 item test which in turn was part of the 50 item test. Half of the items on each test were reading comprehension; and the other half were a mixture of analogy, antonym, and sentence completion items. The means, standard deviations, and KR20 reliabilities of the various tests are shown in Table 1.[1]

Thus, the distributionsof scores for a population of 1586 students on parallel tests of n = 10, 30, and 50 items were available. For each of these n-item tests, four different mastery criteria were considered: c = 50%, 60%, 70%, and 80% items correct. For each combination of the 3 test lengths (n) and the 4 mastery criteria (c), the proportion of the 1586 students consistently classified as master/master or nonmaster/nonmaster on parallel tests of length n was computed, for a total of 3×4 = 12 values of parameter P_c. These parameter values appear in the 12 cells of Tables 2-5 in Appendix A.

For example, in Table 2a when n = 10 items and the mastery criterion is set at c = 50% (or 5 items correct), 67% of the 1586 students were consistently classified on two 10-item parallel tests, i.e., the parameter value is $P_c = .67$. This parameter value of .67 similarly appears in all the tables of Appendix A, as do the parameter values for the other cases considered.

The other numbers in the first cell of Table 2a in Appendix A, are respectively the mean and the standard error of 50 Swaminathan-Hambleton-Algina estimates of parameter value $P_c = .67$, based on 50 random samples of 30 students from the population of 1586 students. For the case of n = 10 items and c = 50%

---

[1] Gary Marco of Educational Testing Service provided the data used in this study, and Barbara Albrecht and Carl Voelz of the University of Wisconsin helped with the analysis. The assistance of each is gratefully acknowledged.

Table 1

Test Statistics[a]

Statistic	Test Form	Test Length		
		10	30	50
Mean	1	4.87	14.49	24.11
	2	4.67	15.18	25.05
Standard Deviation	1	2.00	5.45	8.43
	2	2.07	4.87	7.83
KR20 Reliability	1	.55	.81	.87
	2	.56	.77	.86

[a]Based on a population of 1586 persons.

correct in Table 2a, the mean of 50 Swaminathan-Hambleton-Algina estimates was .68; and the standard error of these estimates was .08, i.e., the estimates tended to deviate from the parameter value of .67 by .08 units on the average. The first cell of Table 2b contains the same type of information for Swaminathan-Hambleton-Algina estimates based on 50 random samples of 300 persons from the population of 1586. Remaining Tables 3-5 of Appendix A provide similar information for the other estimation procedures considered in the study: Marshall-Haertel, Subkoviak, and Huynh.[2]

Results

Swaminathan-Hambleton-Algina Procedure

The Swaminathan et al. (1974, 1975) reliability estimate is simply the proportion of persons in a sample consistently classified as master/master or nonmaster/nonmaster on two tests. As described above, this estimate was computed repeatedly for 50 samples of 30 and for 50 samples of 300 from a population of 1586 persons. The means and standard errors of these estimates, for various test lengths and mastery criteria, are shown in Tables 2a and 2b of Appendix A.

Figure 1a is a graphic representation of Table 2a for samples of 30 persons, while Figure 1b is a graphic representation of Table 2b for samples of 300. In the figures, parameter values are represented by o's; estimate means are represented by ×'s; and standard errors are represented by line intervals (–), indicating the extent to which estimates tend to deviate from the parameter value.

In Figures 1a and 1b, estimate means (×) generally equal corresponding parameter values(0),which suggests that Swaminathan et al. estimates are unbiased,

[2]All computations were done via computer programs written and tested specifically for the project. The interim report of April 30, 1977 describes that phase of the project.

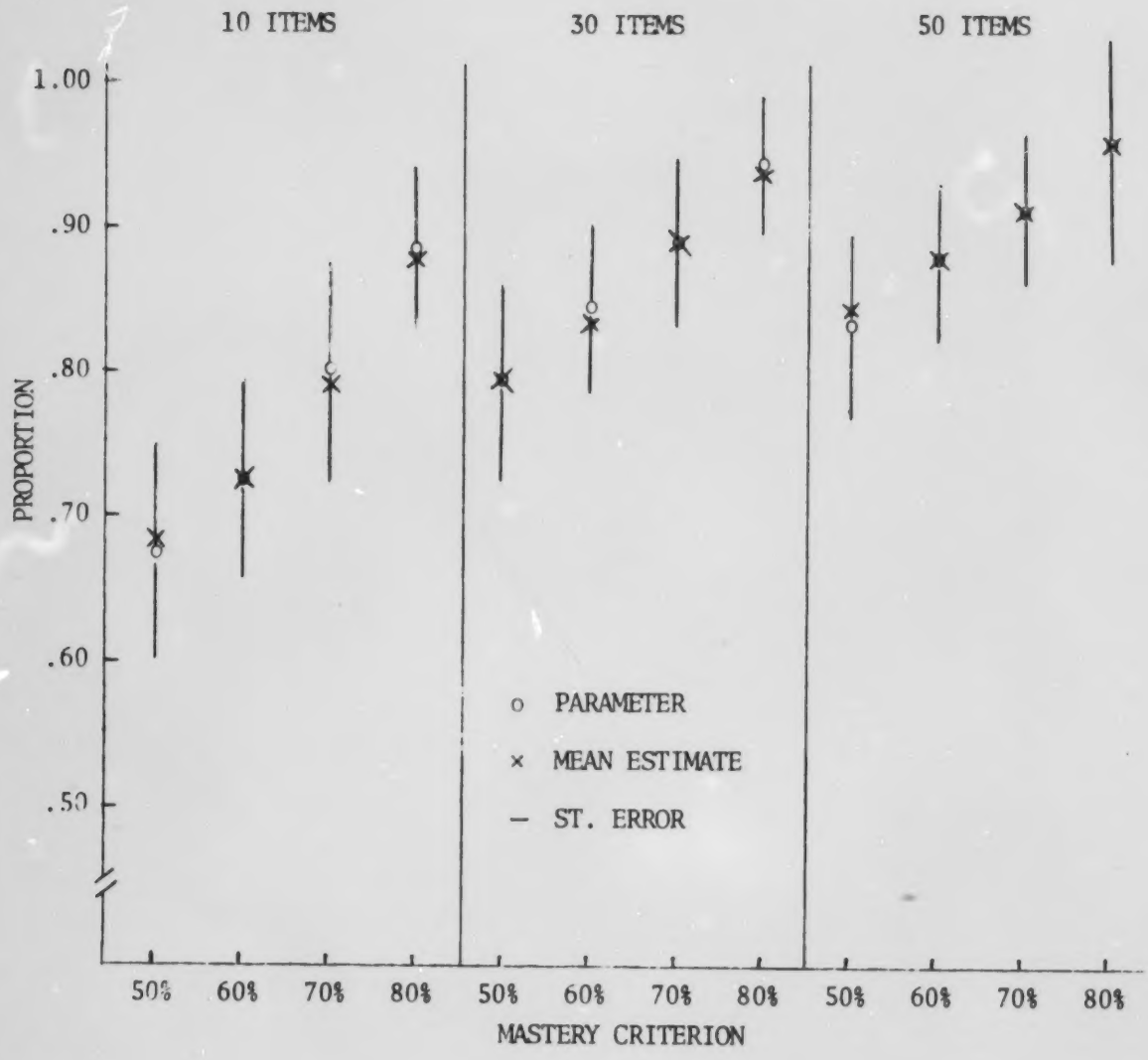


Figure 1a. Means and Standard Errors of Swaminathon-Hambleton-Algina Estimates for Repeated Samples of 30 Persons

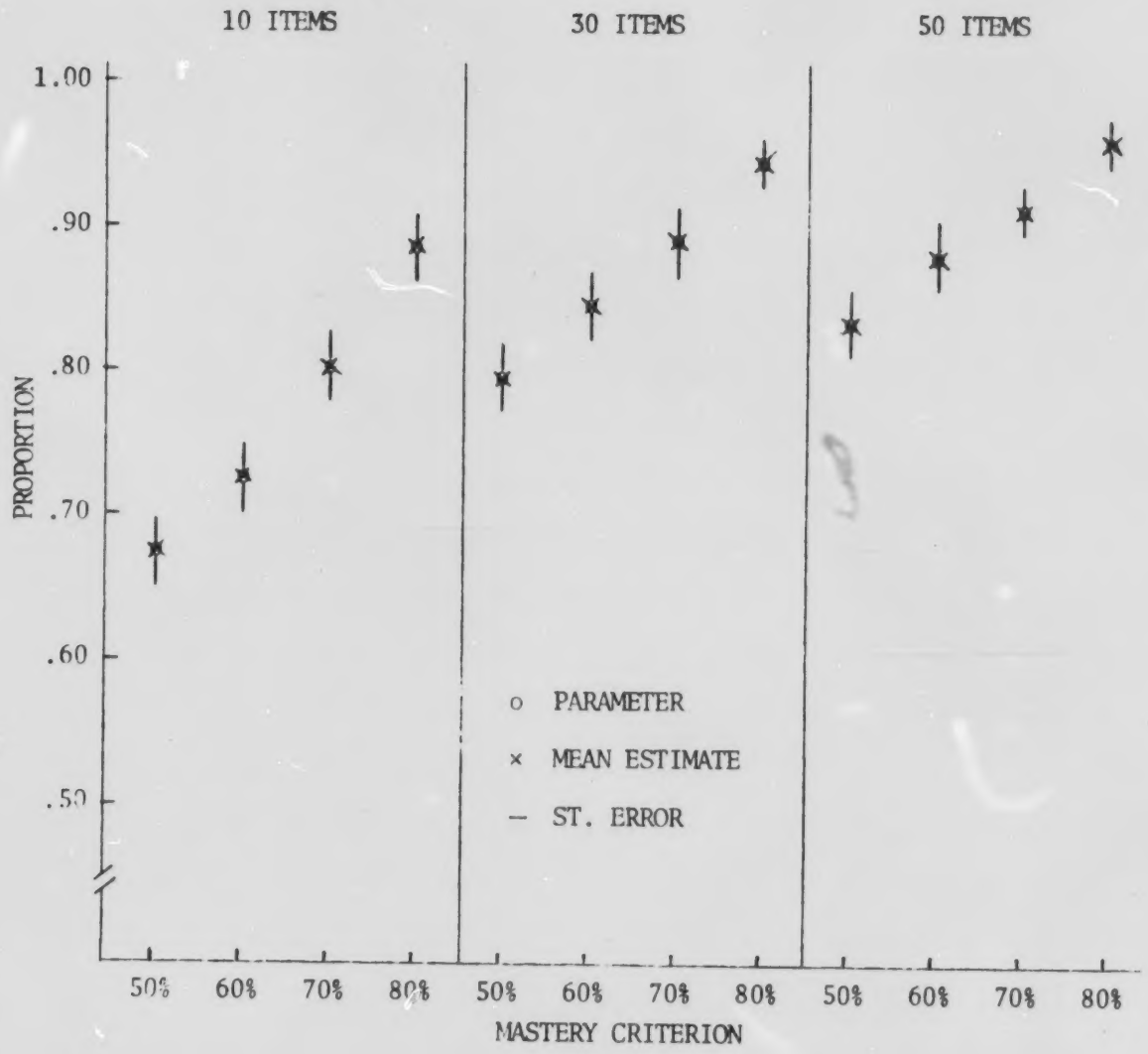


Figure 1b. Means and Standard Errors of Swaminathon-Hambleton-Algina Estimates for Repeated Samples of 300 Persons

as might be expected for this two-test procedure.

However, for classroom size samples of 30 persons or less, the relatively large standard errors of Figure 1a suggest that Swaminathan et al. estimates tend to fluctuate about the parameter value to a greater extent than Marshall-Haertel, Subkoviak, or Huynh estimates (compare the standard errors of Figures 2a, 3a and 4a). Of course, as in Figure 1b, the standard error of estimate can be reduced by increasing sample size to, say, 300 persons, which is not unreasonable for a test publisher or other large-scale user.

It might also be noted in Figures 1a,b or Tables 2a,b that the standard error generally tends to decrease as test length (n) increases and as the mastery criterion (c) increases. These observations are attributable, at least in part, to the fact that as the parametric value of a proportion (P_c) becomes more extreme (large or small), estimates of that proportion tend to be more accurate or less variable. These same trends are repeated in subsequent figures and tables.

It might also be mentioned at this point that the reliability estimates of all subsequent figures and tables are based on a single administration of test Form 1 to samples of 30 and 300 students.[3] This represents a distinct advantage over estimation procedures requiring two test forms or administrations.

Marshall-Haertel Procedure

Procedures that estimate reliability from a single test administration generally substitute certain assumptions for the missing or absent second testing. For example, the Marshall-Haertel procedure makes the hypothetical assumption that if n-item tests were repeatedly administered to an individual student, his or her distribution of observed scores would be binomial, with parameters n (number of items) and p (probability of a correct item response).

---

[3]Estimates based on Form 2 of each test were very similar to those based on Form 1, as indicated in the interim project report of April 30, 1977.

Marshall and Haertel use each student's observed proportion correct score on the actual n-item test to approximate his or her binomial p - parameter, and the group distribution of observed scores on a hypothetical 2n-item test is simulated. This 2n-item test is split into half tests in all possible ways; and an estimate of P_C, the proportion of consistent classification on two tests, is computed for each split. The mean of these various split-half estimates is then taken as the final estimate of P_C. See Marshall and Haertel (1976) for further details.

Figures 2a and 2b are graphic representations of Tables 3a and 3b (Appendix A) for samples of 30 and 300 respectively. Especially for tests of 10 items or less, there appears to be a slight systematic bias in the Marshall-Haertel estimates of Figures 2a,b. The estimate means (×'s) for mastery criteria of 50% and 60%, which are points near the center of the unimodal test score distribution used in the study, tend to overestimate the parameter (o's). Conversely, estimate means (×'s) for mastery criteria of 70% or 80%, which are points in the tails of the distribution, tend to slightly underestimate the parameter (o's). Algina and Noe (1977, p. 6) report the same type of bias in a somewhat different context and relate it to the use of students' observed proportion correct scores as approximations of the binomial p - parameter. The magnitude of such bias should decrease as test length increases, as it does in Figures 2a,b for tests of 30 and 50 items; since observed proportions provide better approximations to the binomial p - parameter as the number of items or trails increases.

Subkoviak Procedure

This procedure assumes that if n-item tests were repeatedly administered to an individual, his or her distribution of observed scores would be compound binomial and that the individual's score on one test does not effect his or her scores on the other tests. Individuals' observed proportion correct scores

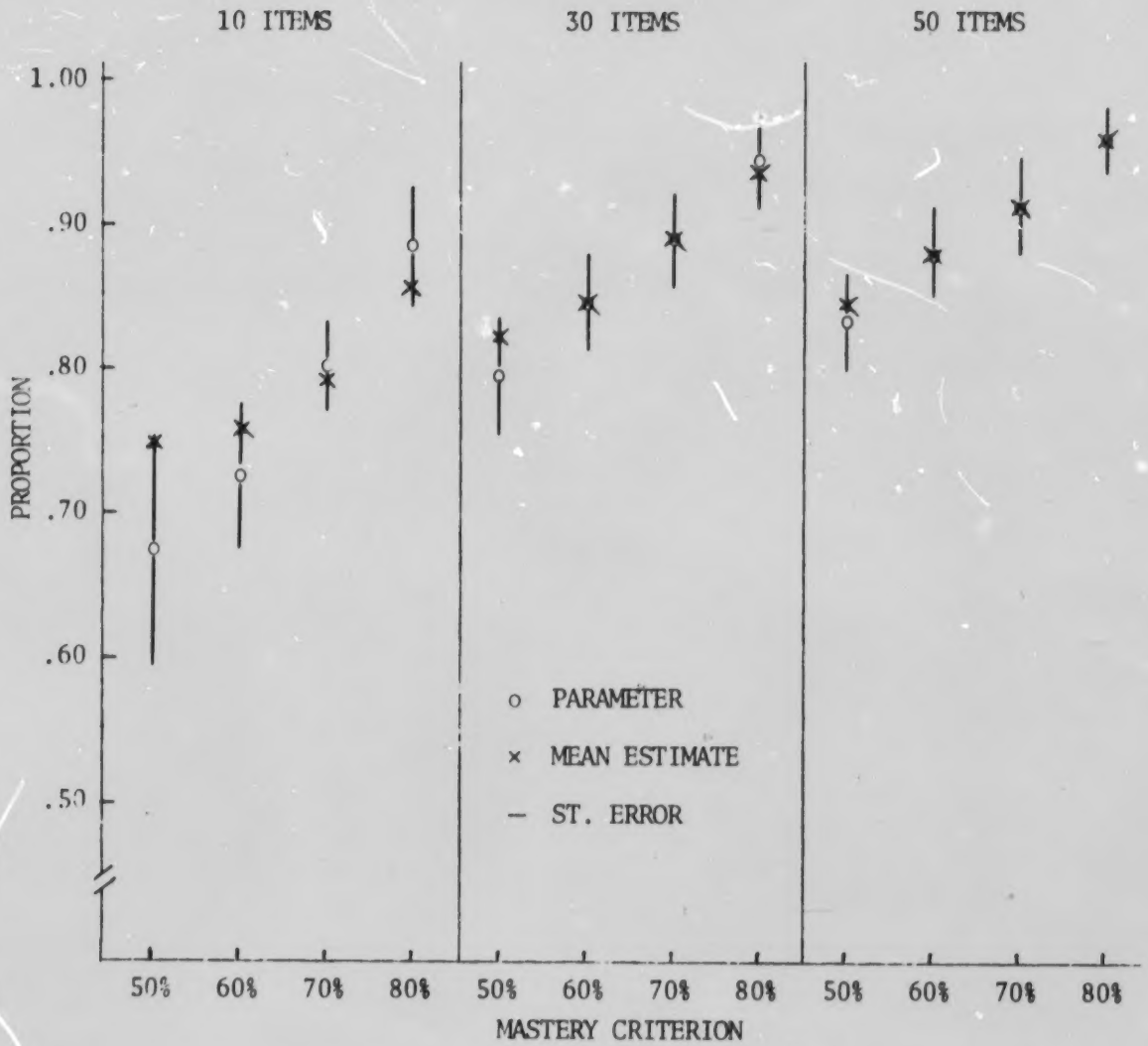


Figure 2a. Means and Standard Errors of Marshall-Haertel Estimates for Repeated Samples of 30 Persons

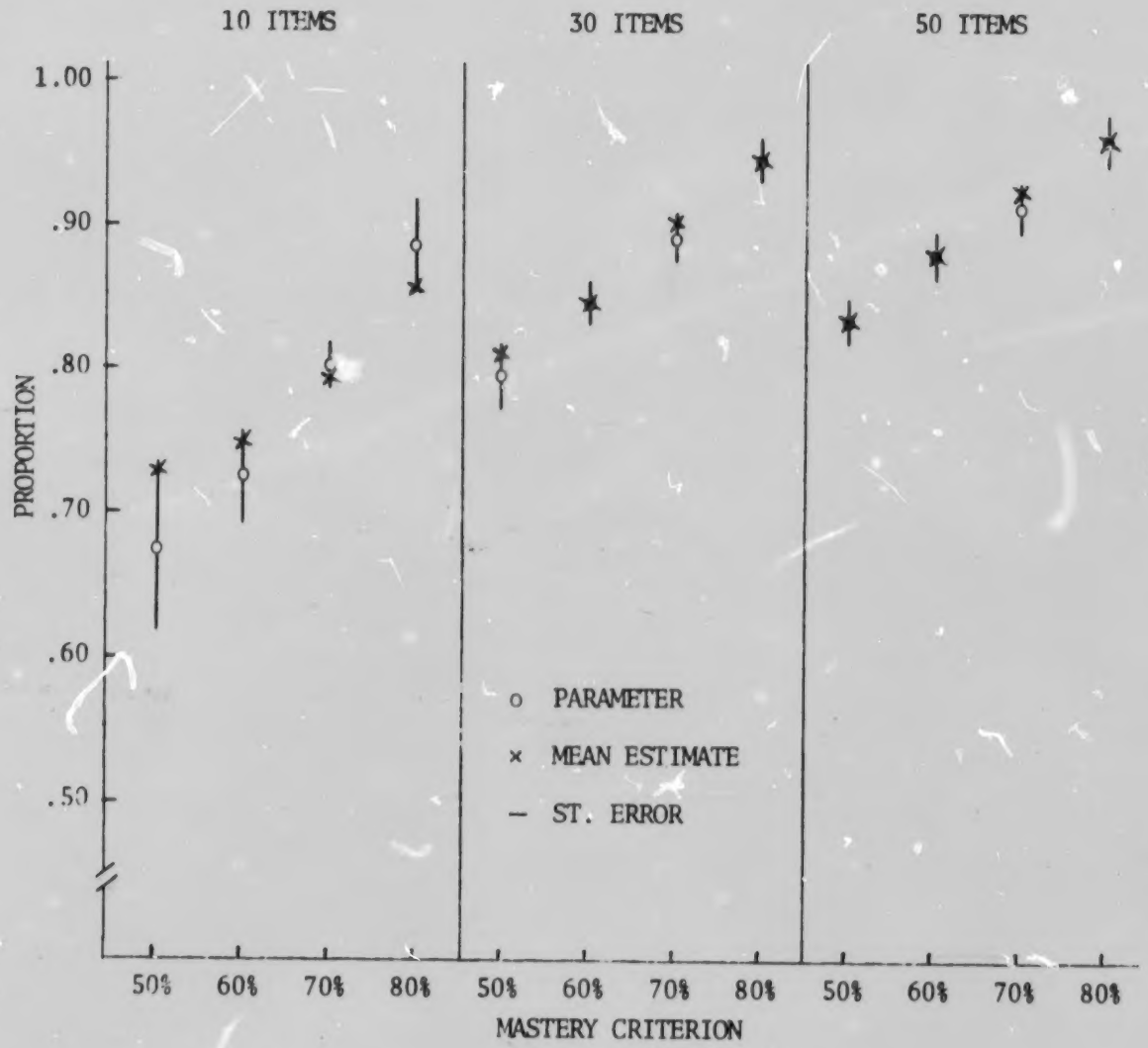


Figure 2b. Means and Standard Errors of Marshall-Haertel Estimates for Repeated Samples of 300 Persons

on a test and the associated KR-20 coefficient are then used to obtain linear regression approximations of individuals' compound binomial p - parameters. A group estimate of P_c, the proportion of consistent classifications on two tests, can then be obtained from the individual compound binomial distributions. See Subkoviak (1976) for details.

Figures 3a,b are graphic representations of Tables 4a,b of Appendix A. Algina and Noe (1977, p. 6) report slight, systematic bias in Subkoviak estimates for simulated data. They found parameter estimates to be too small for mastery criteria near the center of a unimodal test score distribution and too large for mastery criteria in the tails of the distribution. This trend also seems to be present in the 10 item test of Figures 3a,b; however, the evidence of such a trend in the 30 or 50 item tests of Figures 3a,b is somewhat less compelling.

While the Subkoviak procedure procedes reasonably accurate parameter estimates for the unimodal test score distributions considered herein, as evidenced by the relatively small standard errors in Figures 3a,b, grossly inaccurate estimates can occur if linear regression approximations of binomial parameter p are blindly obtained for multimodal data sets (see Huynh, 1977, Counterexample 2; Subkoviak, 1976, p. 269). While more complex regression techniques could be employed to approximate p in such cases, the procedure discussed next provides a more tractable solution for most data sets likely to arise in practice, e.g., unimodal, bimodal, or uniform.

Huynh Procedure

Basically, this procedure assumes that the distribution of observed scores over repeated testing of an individual is binomial with parameters n and p and that such test outcomes are independent of one another. In addition, the distribution of individuals' binomial p - parameters is assumed to be beta in form (see Lavalle,

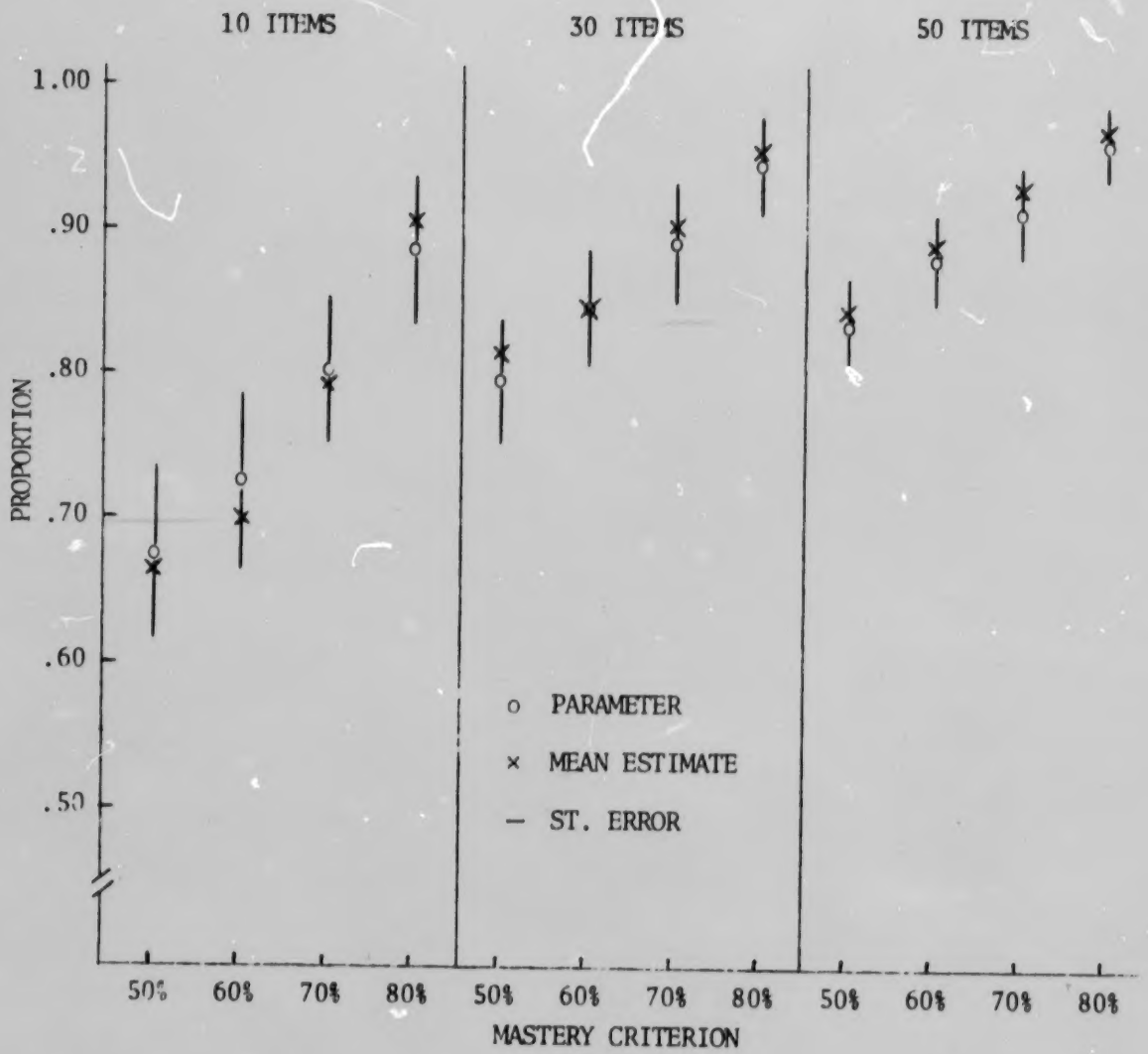


Figure 3a. Means and Standard Errors of Subkoviak Estimates for Repeated Samples of 30 Persons

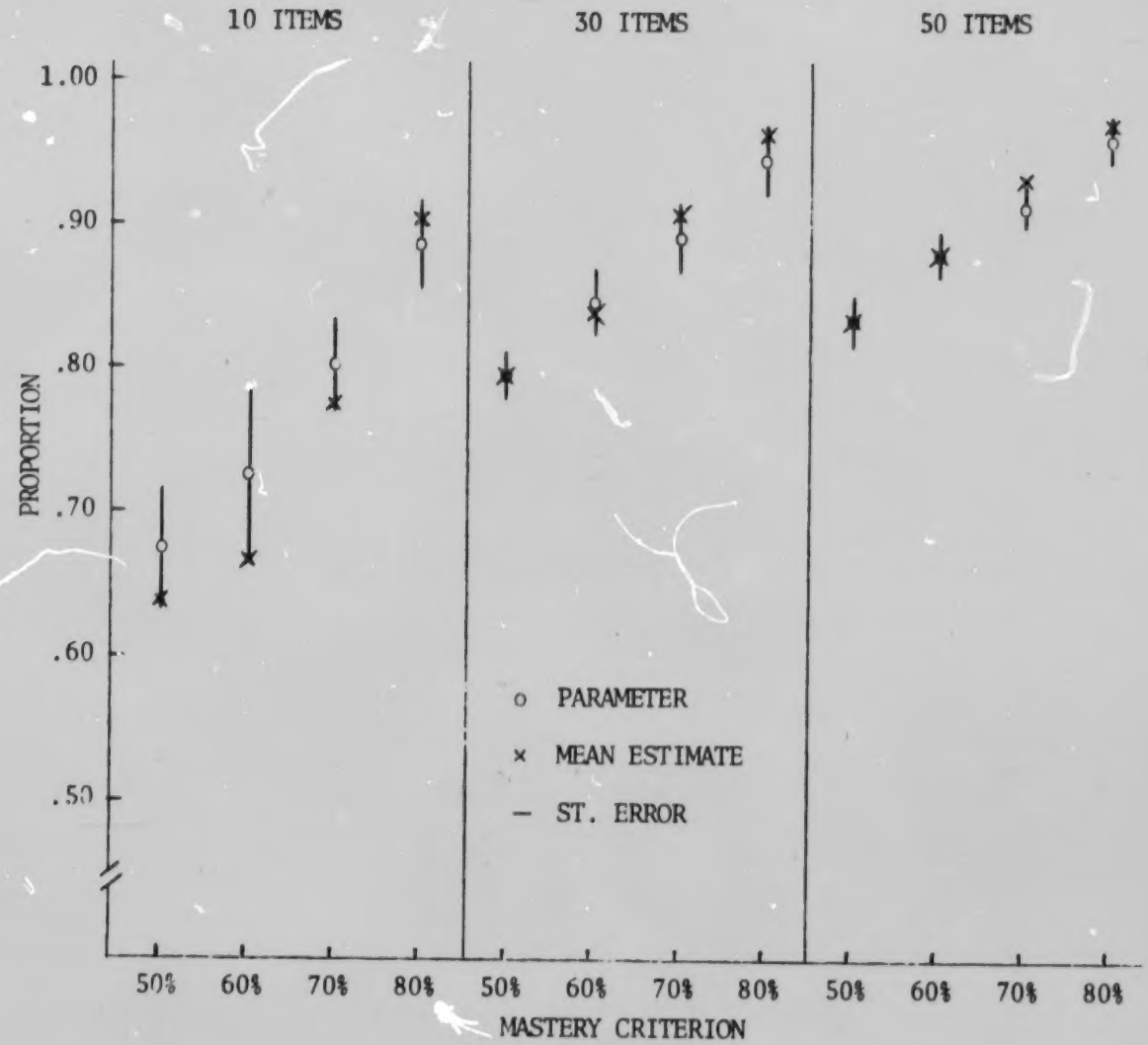


Figure 3b. Means and Standard Errors of Subkoviak Estimates for Repeated Samples of 300 Persons

1970, p. 256 for examples). Under these assumptions, the bivariate distribution of observed scores on two testings for the group is beta-binomial in form and can be simulated from scores on a single test administration. Estimates of parameter P_c can then be obtained from this simulated beta-binomial distribution. See Huynh (1976) for further explication.

Figures 4a,b are graphic representations of Tables 5a,b (Appendix A). The trend in Figures 4a,b would appear to be toward conservative estimation for tests of 10 items or less. However, for tests of 30 or 50 items the Huynh estimates are generally quite good as evidenced by the coincidence of means (×'s) and parameters (o's) as well as the small standard errors of Figures 4a,b.

Conclusions

All four procedures (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976; Swaminathan et al., 1974) appear to provide reasonably accurate estimates of parameter P_c, the proportion of consistent classifications on two mastery tests, for the various cases considered herein. In particular, the Huynh procedure seems especially tractable. The following specific conclusions also appear to be supported by Figures 1-4.

1. The two-test Swaminathan-Hambleton-Algina procedure produces unbiased estimates. However, the standard error of these estimates is relatively large for classroom size samples of 30 or fewer persons. For samples of 300 or more, the standard errors are quite small.
2. For unimodal distributions of scores on tests of 10 items or less, the one-test Marshall-Haertel procedure produces overestimates for mastery criteria near the center of the distribution and underestimates for criteria in the tails. The standard error of Marshall-Haertel estimates is relatively small for classroom size samples of 30 or more.

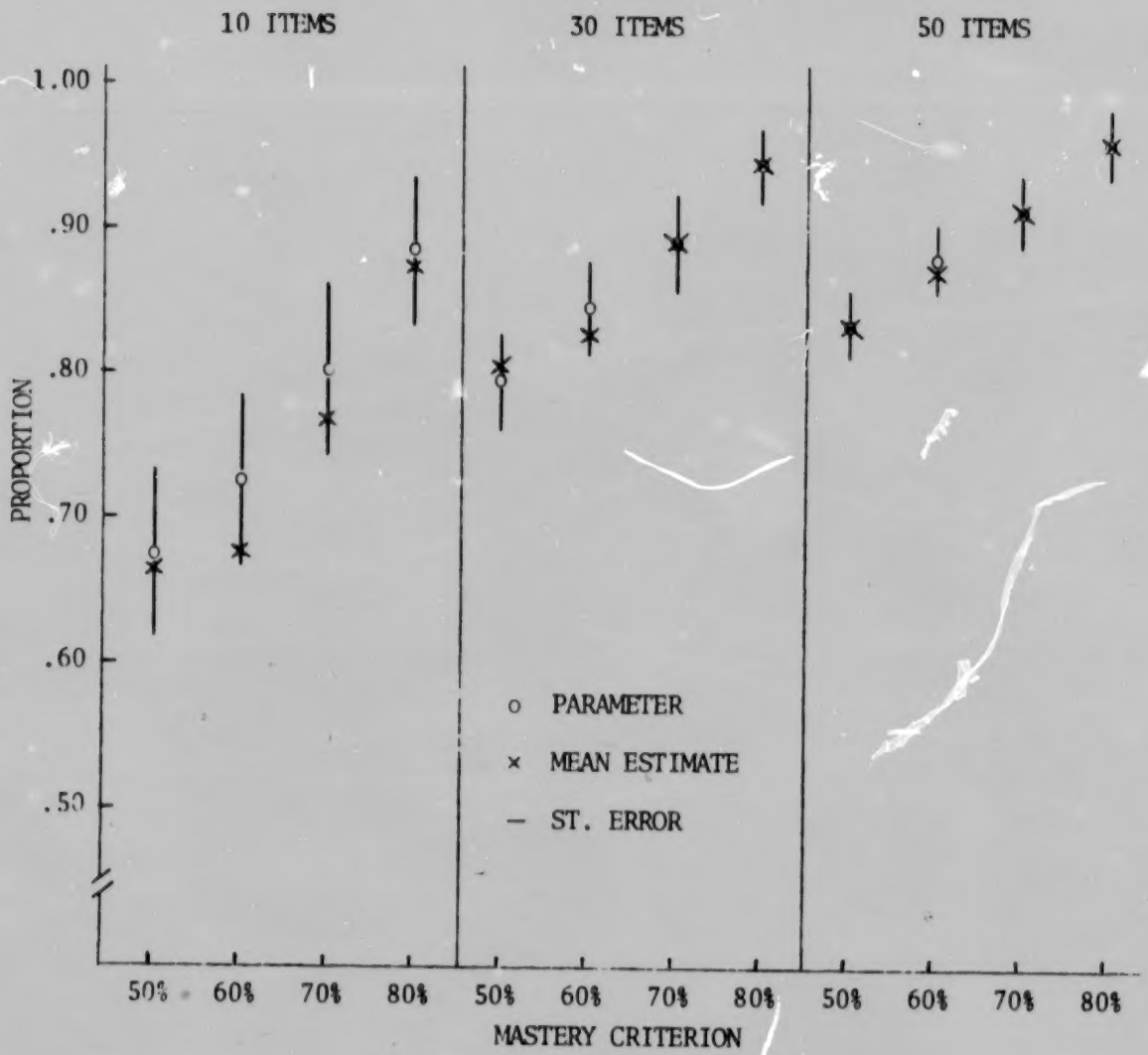


Figure 4a. Means and Standard Errors of Huynh Estimates for Repeated Samples of 30 Persons

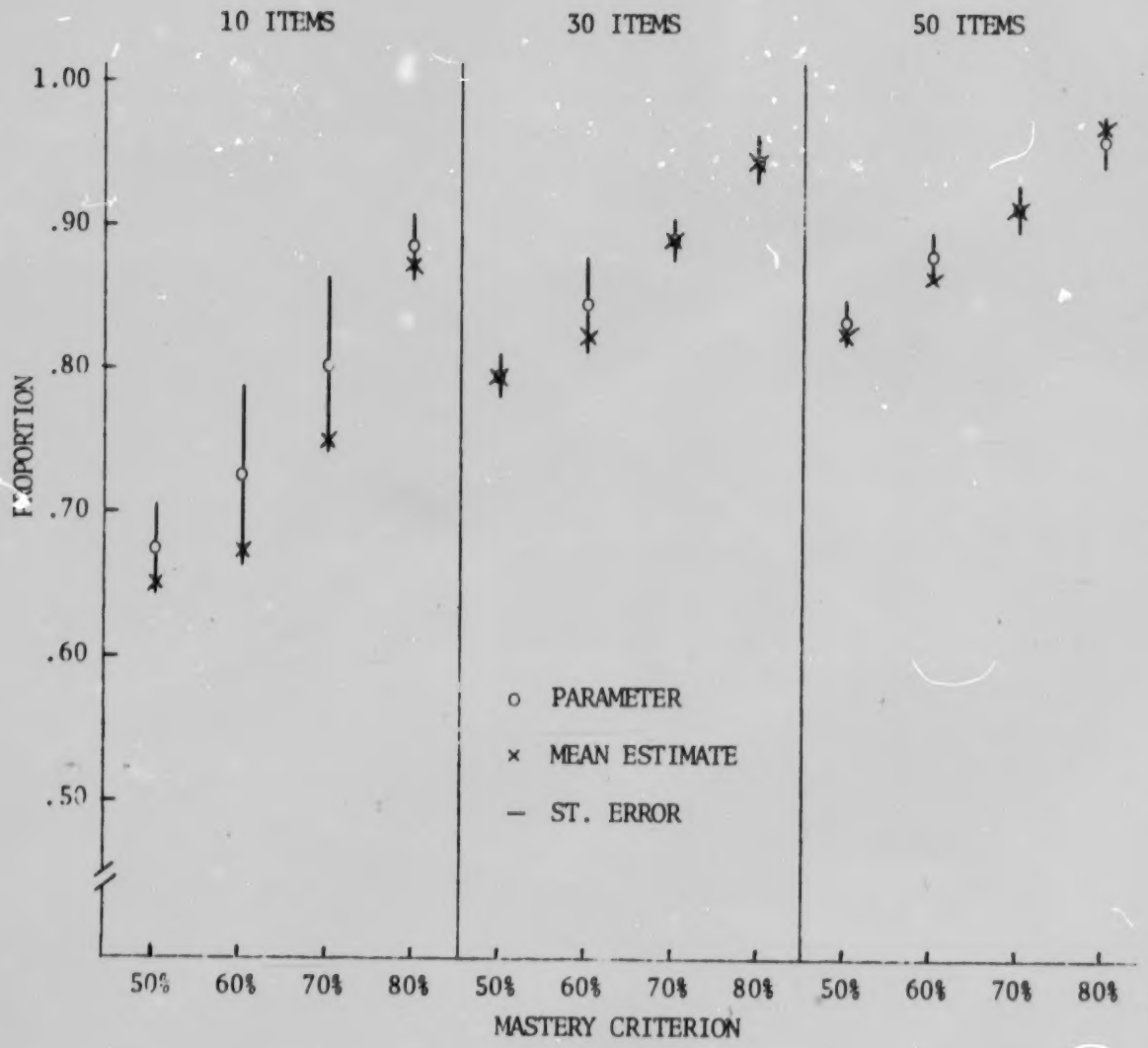


Figure 4b. Means and Standard Errors of Huynh Estimates for Repeated Samples of 300 Persons

3. For unimodal distributions of scores on tests of 10 items or less, the one-test Subkoviak procedure produces underestimates for mastery criteria near the center of the distribution and overestimates for criteria in the tails. The standard errors of Subkoviak estimates is relatively small for samples of 30 or more persons.
4. For unimodal distributions of scores on tests of 10 items or less, the one-test Huynh procedure produces underestimates. The standard errors of Huynh estimates are relatively small for samples of 30 or more persons.

It might also be added that the Huynh procedure appears to have the most sound mathematical basis of the three one-test approaches. The two-test Swaminathan-Hambleton-Algina procedure is also quite tractable in this sense.

Related Research

The foregoing represents the work contracted for in the original project proposal. In addition, a number of papers on related topics were completed and submitted for publication during the grant period. Titles and abstracts of these papers follow. Copies of these papers are also included in Appendix B.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Manuscript submitted for publication, 1977.

Abstract

Four different procedures (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976; Swaminathan, Hambleton & Algina, 1974) have been proposed for estimating the proportion of persons consistently classified as master/master or nonmaster/nonmaster on two mastery tests. Estimates of this proportion were obtained for repeated samples of size N = 30 for each of the above procedures. The estimates were then compared for accuracy to the value of

this proportion in the population of N = 1586 subjects from which the samples were drawn. Both test length and mastery criterion were varied. While reasonably accurate estimates were generally obtained for all four procedures, instances of systematic estimation bias were observed.

Subkoviak, M. J. Further comments on reliability for mastery tests. Manuscript submitted for publication, 1977.

Abstract

This paper illustrates that the various coefficients of classification consistency that have been proposed as measures of reliability for mastery tests have different interpretations and statistical properties. As such, they should not be applied indiscriminately. Rather, a user should employ that coefficient that is most meaningful within the context of a particular problem.

Subkoviak, M. J., & Wilcox, R. Estimating the probability of correct classification in mastery testing. Manuscript submitted for publication, 1977.

Abstract

A procedure is proposed for estimating the proportion of persons in a group that are correctly classified on a mastery test, i.e., the proportion whose observed classification agrees with their true classification. A numerical example is provided, and extensions of the procedure are discussed.

Hubert, L. J., & Subkoviak, M. J. Confirmatory inference and geometric models. Manuscript submitted for publication, 1977.

Abstract

A confirmatory method is discussed for comparing an outside variable to a given geometric model, or alternatively, to the raw data from which the model is derived. The inference procedure is based on relatively simple nonparametric

principles and requires the comparison of a proximity matrix generated from a geometric representation against a second "structure" matrix obtained from the outside variable under study. A number of examples are presented that illustrate how the same statistical approach can be applied in evaluating geometric models that arise in a number of ways, for instance, those produced by some explicit data reduction process, or possibly, models generated by naturally occuring spatial contiguity.

References

Algina, J., & Noe, M.J. An investigation of Subkoviak's single-administration reliability estimate for criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York City, 1977.

Hambleton, R.K., & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Huynh, H. Reliability of criterion-referenced tests: Comments on a paper by Subkoviak. Unpublished manuscript, University of South Carolina, 1977.

La Valle, I.H. An Introduction to Probability, Decision, and Inference. New York: Holt, Rinehart and Winston, 1970.

Marshall, J.L., & Haertel, E.H. The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. Unpublished manuscript, University of Wisconsin, 1976.

Subkoviak, M.J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

Subkoviak, M.J. Further comments on reliability for mastery tests (Laboratory of Experimental Design, Occasional Paper No. 17). Unpublished manuscript, University of Wisconsin, 1977.

Swaminathon, H., Hambleton, R.K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Swaminathon, H., Hambleton, R.K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 87-98.

APPENDIX A

TABLES OF MEANS AND STANDARD ERRORS

Table 2a

Means and Standard Errors of Swaminathon-Hambleton-Algina Estimates for Repeated Samples of 30 Persons

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.68	.79	.84
	St. Error	.08	.07	.06
60%	Parameter	.72	.84	.87
	Mean	.72	.83	.87
	St. Error	.07	.06	.06
70%	Parameter	.80	.88	.91
	Mean	.79	.88	.91
	St. Error	.08	.06	.05
80%	Parameter	.88	.94	.96
	Mean	.87	.93	.96
	St. Error	.06	.05	.08

Table 2b

Means and Standard Errors of Swaminathon-Hambleton-Algina Estimates for Repeated Samples of 300 Persons

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.67	.79	.83
	St. Error	.02	.02	.02
60%	Parameter	.72	.84	.87
	Mean	.72	.84	.87
	St. Error	.02	.02	.02
70%	Parameter	.80	.88	.91
	Mean	.80	.88	.91
	St. Error	.02	.02	.01
80%	Parameter	.88	.94	.96
	Mean	.88	.94	.96
	St. Error	.02	.01	.01

Table 3a

Means and Standard Errors of Marshall-Haertel Estimates for Repeated Samples of 30 Persons

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.74	.82	.84
	St. Error	.08	.04	.03
60%	Parameter	.72	.84	.87
	Mean	.75	.84	.87
	St. Error	.05	.03	.03
70%	Parameter	.80	.88	.91
	Mean	.79	.88	.91
	St. Error	.03	.03	.03
80%	Parameter	.88	.94	.96
	Mean	.85	.93	.96
	St. Error	.04	.03	.02

Table 3b

Means and Standard Errors of Marshall-Haertel Estimates for Repeated Samples of 300 Persons

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.73	.81	.83
	St. Error	.06	.02	.01
60%	Parameter	.72	.84	.87
	Mean	.74	.84	.87
	St. Error	.03	.01	.01
70%	Parameter	.80	.88	.91
	Mean	.79	.89	.92
	St. Error	.01	.01	.01
80%	Parameter	.88	.94	.96
	Mean	.85	.94	.96
	St. Error	.03	.01	.01

Table 4a

Means and Standard Errors of Subkoviak

Estimates for Repeated Samples

of 30 Persons

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.66	.81	.84
	St. Error	.06	.04	.03
60%	Parameter	.72	.84	.87
	Mean	.69	.84	.88
	St. Error	.06	.04	.03
70%	Parameter	.80	.88	.91
	Mean	.79	.89	.93
	St. Error	.05	.04	.03
80%	Parameter	.88	.94	.96
	Mean	.90	.95	.97
	St. Error	.05	.03	.02

Table 4b

Means and Standard Errors of Subkoviak Estimates for Repeated Samples of 300 Persons

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.64	.79	.83
	St. Error	.04	.01	.01
60%	Parameter	.72	.84	.87
	Mean	.66	.83	.87
	St. Error	.06	.02	.01
70%	Parameter	.80	.88	.91
	Mean	.77	.90	.93
	St. Error	.03	.02	.01
80%	Parameter	.88	.94	.96
	Mean	.90	.96	.97
	St. Error	.03	.02	.01

Table 5a

Means and Standard Errors of Huynh

Estimates for Repeated Samples

of 30 Persons

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.66	.80	.83
	St. Error	.06	.03	.02
60%	Parameter	.72	.84	.87
	Mean	.67	.82	.86
	St. Error	.06	.03	.02
70%	Parameter	.80	.88	.91
	Mean	.76	.88	.91
	St. Error	.06	.03	.02
80%	Parameter	.88	.94	.96
	Mean	.86	.94	.96
	St. Error	.05	.02	.02

Table 5b

Means and Standard Errors of Huynh

Estimates for Repeated Samples

of 300 Persons

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.65	.79	.82
	St. Error	.03	.01	.01
60%	Parameter	.72	.84	.87
	Mean	.66	.81	.85
	St. Error	.06	.03	.01
70%	Parameter	.80	.88	.91
	Mean	.74	.88	.91
	St. Error	.06	.01	.01
80%	Parameter	.88	.94	.96
	Mean	.86	.94	.97
	St. Error	.02	.01	.01

APPENDIX B

MANUSCRIPTS PRODUCED DURING THE GRANT PERIOD

Empirical Investigation of Procedures for Estimating Reliability for Mastery Tests

Michael J. Subkoviak

University of Wisconsin

Running head: Reliability for Mastery Tests

Abstract

Four different procedures (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976; Swaminathan, Hambleton & Algina, 1974) have been proposed for estimating the proportion of persons consistently classified as master/master or nonmaster/nonmaster on two mastery tests. Estimates of this proportion were obtained for repeated samples of size $\underline{N}$ = 30 for each of the above procedures. The estimates were then compared for accuracy to the value of this proportion in the population of $\underline{N}$ = 1586 subjects from which the samples were drawn. Both test length and mastery criterion were varied. While reasonably accurate estimates were generally obtained for all four procedures, instances of systematic estimation bias were observed.

Empirical Investigation of Procedures for Estimating Reliability for Mastery Tests

For present purposes, a mastery test can be loosely defined as a test with a single cutting score, c, that determines mastery and nonmastery classes -- scores above and below c respectively. In this context, reliability refers to the consistency of mastery-nonmastery decisions over repeated test administrations (Hambleton & Novick, 1973, pp. 166-167). Accordingly, the proportion of consistent mastery/mastery and nonmastery/nonmastery classifications on *two* tests with cutting score c, symbolized P_c, has been proposed as a raw index of reliability in this context (Swaminathan, Hambleton & Algina, 1974, 1975). In addition, three procedures for estimating proportion P_c from scores on a *single* test have emerged (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976). Thus, a teacher or other test user is faced with the problem of choosing among four different procedures for estimating P_c in the absence of any clear guidelines. The purpose of this brief note is to report the results of a simple empirical exercise in which population values of P_c were compared to sample estimates of P_c. Specifically, the proportion of consistent classifications on two tests for a population of 1586, a P_c - parameter, was compared for accuracy to four different P_c - estimates (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976; Swaminathan et al., 1974) for repeated samples of 30 from the same population. The results thus illustrate the extent and nature of discrepancies between parameter and estimates. The results also have indirect relevance for the process of estimating coefficient kappa, a function of P_c that has also been proposed as an index of reliability for mastery tests (see Huynh, 1976; Subkoviak, 1977; Swaminathan et al., 1974).

Method and Results

The data base consisted of the responses of 1586 students to parallel forms of 10, 30, and 50 items each from the Scholastic Aptitude Test. (The 10-item forms were included as part of the 30 item forms which in turn were part of the 50 item forms.) The means, standard deviations, and KR20 reliabilities of the various forms are shown in Table 1. Half of the items on each form were reading comprehension; and the other half were a mixture of analogy, antonym, and sentence completion items.

Insert Table 1 here

Thus, the distribution of scores for 1586 students on parallel tests of n = 10, 30, and 50 items were available. Four different mastery criteria were considered for each n-item test: c = 50%, 60%, 70%, and 80% correct. For each combination of the three test lengths (n) and the four mastery criteria (c), the proportion (P_c) of the 1586 students consistently classified as master/master or nonmaster/nonmaster (on parallel forms of length n) was computed, for a total of 12 values of parameter P_c. These values appear in each cell of Tables 2-5. For example, when n = 10 and the mastery criterion is set at c = 50% (or 5 items) correct in Tables 2-5, 67% of the 1586 students were consistently classified on two 10-item parallel forms, i.e., P_c = .67. The other numbers in each cell of Tables 2-5 are the mean and the standard error of 50 estimates of parameter P_c based on 50 random samples of 30 students from the same population of 1586. For example, when n = 10 and c = 50% in Table 2, the mean of the 50 estimates is .68; and their standard error is .08, i.e., the estimates tend to deviate from the parameter value (.67) by .08 units on the average.

Swaminathan-Hambleton-Algina Procedure

In Table 2, the fact that cell means generally equal corresponding parameter values (and the two do not deviate in any obvious systematic fashion) suggests that Swaminathan et al. estimates are unbiased, as might be expected for this two test procedure. However, as will become apparent, the standard errors of Table 2 tend to be somewhat larger than those of Tables 3-5, suggesting that Swaminathan et al. estimates tend to fluctuate about the parameter value to a greater extent than Marshall-Haertel, Huynh, or Subkoviak estimates. Of course, the standard error could easily be reduced by increasing the sample size from $N = 30$ to, say, $N = 100$.

Insert Table 2 here

It might also be noted in Table 2 that the standard error generally tends to decrease as test length (n) increases and as the mastery criterion (c) increases. These observations are attributable, at least in part, to the fact that as the parametric value of a proportion (P_c) becomes more extreme (large or small), estimates of that proportion tend to be more accurate (less variable). These same trends are repeated in subsequent Tables 3-5. It should also be mentioned at this point that the reliability estimates of Tables 3-5 are based on one administration of test Form 1 to samples of 30 students. Estimates based on parallel Form 2 of each test are very similar and are not reported here.

Marshall-Haertel Procedure

This procedure assumes that distribution of observed scores over repeated testing of a fixed, individual student is binomial in form. Students' observed proportion correct scores on an actual n-item test are used as approximations to the binomial p - parameter, and the group distribution of observed scores on a

hypothetical 2n-item test is simulated. This 2n-item test is split into half-tests in all possible ways, and an estimate of P_C is computed for each split. The mean of these various split-half estimates is then taken as the final estimate of P_C. See Marshall and Haertel (1976) for further details.

Table 3 contrasts Marshall-Haertel means with parameter values.

Insert Table 3 here

There appears to be a slight, systematic bias in the estimates (means) of Table 3. The estimates in the top two rows, corresponding to mastery criteria near the mean of the test score distribution, tend to overestimate the parameter, especially for short tests. Conversely, estimates in the bottom two rows, corresponding to mastery criteria in the tails of the distribution, tend to slightly underestimate the parameter. Algina and Noe (1977, p. 6) report the same type of bias in a somewhat different context and relate it to the use of students' observed proportion correct scores as approximations of the binomial p - parameter. The magnitude of such bias should decrease as test length (n) increases (as in Table 3); since observed proportions provide better approximations to the binomial p - parameter as the number of items (trials) increases.

Subkoviak Procedure

This procedure assumes that the distribution of observed scores over repeated testing is compound binomial and that test outcomes are independent for a fixed individual. Observed proportion correct scores on a test and the associated KR-20 coefficient are used to obtain linear regression approximations to the compound binomial p - parameter. A group estimate of P_C, the proportion of consistent classifications on two tests, can then be obtained from the individual compound binomial distributions. See Subkoviak (1976) for details.

Table 4 contrasts Subkoviak means with parameter values. Algina and Noe

Insert Table 4 here

(1977, p. 6) report slight, systematic bias in Subkoviak estimates (based on linear regression approximations of binomial parameter p) for simulated data. Specifically, they found estimates of P_c to be too small for mastery criteria (c) near the test score distribution mean and too large for mastery criteria in the tails of the distribution. Such a trend is not obvious in Table 4. If, indeed, there is a trend, it may be toward slight underestimation for short tests ($n = 10$) and slight overestimation for long tests ($n = 50$); but the evidence is not totally compelling.

While the Subkoviak procedure produces reasonably accurate P_c-estimates in this case, as evidenced by the small standard errors, grossly inaccurate estimates can occur if linear regression approximations of binomial parameter p are blindly obtained for multimodal data sets (see Huynh, 1977, Counterexample 2; Subkoviak, 1976, p. 269). While more complex regression techniques could be employed to approximate p in such cases, the procedure discussed next provides a more tractable solution for most data sets likely to arise in practice, e.g., unimodal, uniform, or bimodal.

Huynh Procedure

Basically, this procedure assumes that the distribution of observed scores over repeated testing is binomial and that test outcomes are independent for a fixed individual. In addition, the associated distribution of individuals' p - parameters is assumed to be beta in form (see LaValle, 1970, p. 256 for examples). Under these assumptions, the bivariate distribution of scores on two testings for the group is beta-binomial in form and can be approximated from scores on a single test administration. Estimates of P_c can then be obtained from the simulated beta-binomial distribution. See

Huynh (1976) for further explication.

Insert Table 5 here

Huynh means and parameter values are contrasted in Table 5. The trend in Table 5 (as in Table 4) would appear to be toward conservative estimation for short tests. However, the estimates are generally quite good as evidenced by the small standard errors.

Conclusions

All four procedures (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976; Swaminathan et al., 1974) appear to provide reasonably accurate estimates of $\underline{P}_{\underline{C}}$, the proportion of consistent classifications on two mastery tests, for the various cases considered. In particular, the Huynh procedure seems especially tractable. Table 6 shows the means and standard errors of Huynh estimates based on 50 samples of 300 persons from the same population of 1586. Test publishers that employ large pilot samples might expect results like these.

Insert Table 6 here

While it might seem inappropriate to employ SAT data in the present study rather than mastery test data, this would not appear to be a serious limitation. Statistically speaking, the key issue is the performance of each estimation procedure as the parametric value of proportion $\underline{P}_{\underline{C}}$ ranges between .50 and 1.00, regardless of the data base. In fact, it is interesting to note that the Marshall-Haertel, Subkoviak, and Huynh procedures, which basically assume item homogeneity, produced accurate estimates for the heterogeneous SAT items employed herein.

References

Algina, J., & Noe, M. J. An investigation of Subkoviak's single-administration reliability estimate for criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York City, 1977.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Huynh, H. Reliability of criterion-referenced tests: Comments on a paper by Subkoviak. Unpublished manuscript, University of South Carolina, 1977.

La Valle, I. H. An Introduction to Probability, Decision, and Inference. New York: Holt, Rinehart and Winston, 1970.

Marhsall, J. L., & Haertel, E. H. The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. Unpublished manuscript, University of Wisconsin, 1976.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

Subkoviak, M. J. Further comments on reliability for mastery tests (Laboratory of Experimental Design, Occasional Paper No. 17). Unpublished manuscript, University of Wisconsin, 1977.

Swaminathon, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Swaminathon, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 87-98.

Footnotes

This research was made possible by Grant No. NIE-G-76-0088 from the National Institute of Education.

Gary Marco of the College Examination Entrance Board and Educational Testing Service provided the data used in this study, while Barbara Albrecht and Carl Voelz helped with the analyses. The assistance of each is gratefully acknowledged.

Table 1

Test Statistics[a]

Statistic	Test Form	Test Length		
		10	30	50
Mean	1	4.87	14.49	24.11
	2	4.67	15.18	25.05
Standard Deviation	1	2.00	5.45	8.43
	2	2.07	4.87	7.83
KR20 Reliability	1	.55	.81	.87
	2	.56	.77	.86

[a]N = 1586

Table 2

Means and Standard Errors of Swaminathon-

Hambleton- Algina Estimates[a]

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.68	.79	.84
	St. Error	.08	.07	.06
60%	Parameter	.72	.84	.87
	Mean	.72	.83	.87
	St. Error	.07	.06	.06
70%	Parameter	.80	.88	.91
	Mean	.79	.88	.91
	St. Error	.08	.06	.05
80%	Parameter	.88	.94	.96
	Mean	.87	.93	.96
	St. Error	.06	.05	.08

[a]Means and standard errors are based on 50 samples of 30 persons.

Table 3

Means and Standard Errors of Marshall-Haertel Estimates[a]

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.74	.82	.84
	St. Error	.08	.04	.03
60%	Parameter	.72	.84	.87
	Mean	.75	.84	.87
	St. Error	.05	.03	.03
70%	Parameter	.80	.88	.91
	Mean	.79	.88	.91
	St. Error	.03	.03	.03
80%	Parameter	.88	.94	.96
	Mean	.85	.93	.96
	St. Error	.04	.03	.02

[a]Means and standard errors are based on 50 samples of 30 persons.

Table 4

Means and Standard Errors of Subkoviak Estimates[a]

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.66	.81	.84
	St. Error	.06	.04	.03
60%	Parameter	.72	.84	.87
	Mean	.69	.84	.88
	St. Error	.06	.04	.03
70%	Parameter	.80	.88	.91
	Mean	.79	.89	.93
	St. Error	.05	.04	.03
80%	Parameter	.88	.94	.96
	Mean	.90	.95	.97
	St. Error	.05	.03	.02

[a]Means and standard errors are based on 50 samples of 30 persons.

Table 5

Means and Standard Errors of Huynh Estimates[a]

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.66	.80	.83
	St. Error	.06	.03	.02
60%	Parameter	.72	.84	.87
	Mean	.67	.82	.86
	St. Error	.06	.03	.02
70%	Parameter	.80	.88	.91
	Mean	.76	.88	.91
	St. Error	.06	.03	.02
80%	Parameter	.88	.94	.96
	Mean	.86	.94	.96
	St. Error	.05	.02	.02

[a]Means and standard errors are based on 50 samples of 30 persons.

Table 6

Means and Standard Errors of Huynh Estimates[a]

Mastery Criterion (c)	Statistical Index	Test Length (n)		
		10	30	50
50%	Parameter	.67	.79	.83
	Mean	.65	.79	.82
	St. Error	.03	.01	.01
60%	Parameter	.72	.84	.87
	Mean	.66	.81	.85
	St. Error	.06	.03	.01
70%	Parameter	.80	.88	.91
	Mean	.74	.88	.91
	St. Error	.06	.01	.01
80%	Parameter	.88	.94	.96
	Mean	.86	.94	.97
	St. Error	.02	.01	.01

[a]Means and standard errors are based on 50 samples of 300 persons.

Estimating the Probability of Correct Classification in Mastery Testing

Michael J. Subkoviak
University of Wisconsin

Rand Wilcox
University of California at Los Angeles

Running head: Probability of Correct Classification in Mastery Testing

Abstract

A procedure is proposed for estimating the proportion of persons in a group that are correctly classified on a mastery test, i.e., the proportion whose observed classification agrees with their true classification. A numerical example is provided, and extensions of the procedure are discussed.

Estimating the Probability of Correct Classification in Mastery Testing

Keats and Lord (1962) have proposed a relatively simple mathematical model for test scores that has a number of practical applications. For instance, Huynh (1976) and Subkoviak and Albrecht (1977) have demonstrated empirically that the model can be used to estimate the degree to which persons are consistently classified as masters or nonmasters on parallel mastery tests, where a pass-fail test with a cut-off score of 75% correct is one example of a mastery test.

The purpose of the present paper is to illustrate that the Keats-Lord model is also useful for estimating the extent to which persons are correctly classified, i.e., for estimating the proportion of persons whose classification based on observed score agrees with their classification based on true score. In addition, extensions of the procedure to the case of polychotomous classification and further generalizations of the Keats-Lord model are noted.

The Keats-Lord Model

Let us begin with the following notational definitions:

n = number of test items;

x = number of items correctly answered by an individual;

p = unknown true proportion correct score for an individual, i.e., the mean of an individual's observed proportion correct scores (x/n) over repeated parallel tests;

π = mastery criterion expressed as a proportion, e.g., $p \geq \pi$ implies true mastery and $p < \pi$ implies true nonmastery;

c = mastery criterion expressed as the number of items correct on an n-item test, e.g., $x \geq c$ implies observed mastery and $x < c$ implies observed nonmastery; c equals the smallest integer greater than or equal to $n\pi$.

Of interest here is the extent to which persons' observed mastery-nonmastery classifications (i.e., $x \geq c$ or $x < c$) are correct or, in other words, agree with their true mastery-nonmastery states (i.e., $p \geq \pi$ or $p < \pi$). One natural index of agreement in this sense is the probability that the observed classification corresponds to the true classification for a typical examinee. This probability will be symbolized Q, and thus Q represents the probability of correct classification for an examinee randomly selected from the population of potential examinees. Mathematically Q can be expressed as:

$$Q = P(x < c, p < \pi) + P(x \geq c, p \geq \pi), \tag{1}$$

where $P(x < c, p < \pi)$ and $P(x \geq c, p \geq \pi)$ are respectively the probability of correct nonmastery and correct mastery classifications.

Under the assumptions of the Keats-Lord model (discussed below), it can be shown (see the Appendix) that Q in Equation 1, the probability of a correct classification, is given by:

$$Q = \{1/B(\ell+1, m+1)\}\{\sum_{x=0}^{c-1}\binom{n}{x} B(\ell+x+1, m+n-x+1)\ I_{\pi}(\ell+x+1, m+n-x+1) + \sum_{x=c}^{n}\binom{n}{x} B(\ell+x+1, m+n-x+1)[1-I_{\pi}(\ell+x+1, m+n-x+1)]\} \tag{2}$$

If the mean (u), variance (σ^2), and Kuder-Richardson 21 reliability coefficient (ρ) of observed x-scores are estimated from a reasonably large sample of testees, the various terms in Equation 2 can be evaluated as follows:

n = number of test items;

π = mastery criterion expressed as a proportion;

$\underline{c}$ = smallest integer greater than or equal to $\underline{n}\pi$;

$\rho = [n/(n-1)][1-u(n-u)/(n\sigma^2)]$;

$\underline{\ell} = u(1/\rho-1)-1$;

$\underline{m} = (\underline{n}-u)(1/\rho-1)-1$;

$\binom{\underline{n}}{\underline{x}} = \underline{n}!/[(\underline{n}-\underline{x})!\underline{x}!]$;

$\underline{B}(\ ,\ )$ = beta function which can be calculated by common computer routines or can be found in standard mathematical tables;

$\underline{I}_{\pi}(\ ,\ )$ = incomplete beta function which can also be computed by common computer routines or can be found in standard mathematical tables.[1]

The Keats-Lord model upon which Equation 2 is based involves two basic assumptions about the nature of true and error scores. It should be noted in passing that the model and its assumptions have been shown to fit a variety of real data sets quite adequately (Keats & Lord, 1962). First, it is assumed that the distribution of true scores ($\underline{p}$) for the population of examinees is some member of the beta family of distributions. This family includes distributions having the usual bell-shape, as well as rectangular and U-shapes (see LaValle, 1970, p. 256 for more examples). As such, the beta assumption is quite liberal in that it accomodates a wide range of possible true score distributions.

Second, if n-item tests were repeatedly administered to a single individual with true score $\underline{p}$, it is assumed that his or her distribution of observed scores ($\underline{x}$) would be binomial with parameters $\underline{n}$ and $\underline{p}$. This assumption tends to follow, for instance, if items are scored 0 and 1; if the outcome on one item does not affect the outcome on others; and if items are equally difficult. While the latter two conditions rarely, if ever, occur in practice, the quality of empirical results reported by Keats and Lord (1962) seems to indicate that the model is robust with respect to such violations.

Example

Suppose an n=4 item test was administered to N=100 students; and suppose the test scores $x = 0,1,2,3$ and 4 occurred with frequencies $f(x) = 8,25,34,25$ and 8 respectively. The sample mean and variance are $\hat{u} = \Sigma x\, f(x)/N = 2.00$ and $\hat{\sigma}^2 = \Sigma(x-\hat{u})^2 f(x)/(N-1) = 1.15$; and the estimate of KR21 is $\hat{\rho} = [n/(n-1)] \times [1-\hat{u}(n-\hat{u})/(n\hat{\sigma}^2)] = .17$. The quantities ℓ and m required by Equation 2 are equal in this particular instance, i.e., $\hat{\ell} = \hat{u}(1/\hat{\rho}-1)-1 = 8.76$ and $\hat{m} = (n-\hat{u})(1/\hat{\rho}-1)-1 = 8.76$; but this is not true in general. Now if the mastery criterion is set at, say, $\pi = .85$, then $c = 4$ in Equation 2; since 4 is the smallest integer greater than or equal to $n\pi = 3.40$.

Summarizing, the parameters required by Equation 2 are: n=4, $\hat{\ell}$=$\hat{m}$=8.76, π=.85 and c=4. Thus by Equation 2 the probability of a correct classification is:

$$Q = \{1/B(9.76,\ 9.76)\}\{\sum_{x=0}^{3}\binom{4}{x}B(9.76+x,\ 13.76-x)I_{.85}(9.76+x,\ 13.76-x) + \sum_{x=4}^{4}\binom{4}{x}B(9.76+x,\ 13.76-x)[1-I_{.85}(9.76+x,\ 13.76-x)]\}$$

where $\binom{4}{x}$, $B(\ ,\ )$ and $I_{.85}(\ ,\ )$ can be obtained via computer or standard mathematical tables (e.g., Pearson, 1956) and are as shown in Table 1. Substituting

Insert Table 1 about here

these values into the above equation gives the following result:

$$Q = \{1/(.15\times10^{-5})\}\{(1)(.12\times10^{-6})(.99)+(4)(.95\times10^{-7})(.99)+(6)(.87\times10^{-7})(.99)+(4)(.95\times10^{-7})(.99)+(1)(.12\times10^{-6})(1-.99)\}$$

$$= .93.$$

Thus, it is likely that 93 out of the 100 students tested are correctly classified.

Discussion

In the example above, $\underline{Q}=.93$ when the mastery criterion is set at $\pi=.85$. However, larger or smaller values of $\underline{Q}$ could be obtained for this same data set by simply changing π. In fact, the magnitude of $\underline{Q}$ is affected by a number of such factors. Specifically, the probability of correct classification, $\underline{Q}$, tends to increase: (a) as the density of true scores about the mastery criterion decreases and (b) as the number of test items increases. Thus, $\underline{Q}$ would assume larger values for settings of the π-criterion in the tails of a bell-shaped distribution (as in the example above) than for settings in the middle of the distribution. Secondly, for a particular setting of π such as .85, $\underline{Q}$ would be larger for a $2\underline{n}$-item test than for an $\underline{n}$-item test.

In regard to generalization, it is worth noting that the probability of correct classification can also be estimated for the case of three or more levels of mastery. For example, let π_1 and π_2 (where $\pi_1 < \pi_2$) represent different degrees of mastery on the true score scale; and let $\underline{c}_1$ and $\underline{c}_2$ be the corresponding criteria on the observed score scale. Three categories of mastery are thus possible, and Equation 1 can be generalized as follows:

$$\underline{Q} = \underline{P}(\underline{x}<\underline{c}_1,\ \underline{p}<\pi_1) + \underline{P}(\underline{c}_1\leq\underline{x}<\underline{c}_2,\ \pi_1\leq\underline{p}<\pi_2) + \underline{P}(\underline{c}_2\leq\underline{x},\ \pi_2\leq\underline{p}). \qquad (3)$$

The argument outlined in the Appendix can then be applied to Equation 3 to obtain an expression similar to Equation 2.

Extensions of the Keats-Lord model, on which Equation 2 is based, are also possible. For example, Lord (1965) has suggested a slightly more complex model that involves somewhat more general and reasonable assumptions regarding true and error score distributions (see also Lord, 1969). However, a Monte Carlo study by Wilcox (1977) seems to suggest that the more complex model adds little in the way of accuracy to probability estimates, which are of interest here. Of course this does not imply that the more complex model lacks utility for

certain other purposes, e.g., for simulating or estimating complete bivariate distributions of true and observed scores.

Appendix

The beta assumption of the Keats-Lord model implies that a true score, p, can assume any value within the interval 0 to 1; while the binomial assumption implies that an observed score, x, takes on only integral values between 0 and n. Thus, Equation 1 can be written as a double summation of $P(x,p)$ values, where $P(x,p)$ represents the joint probability distribution of examinees' observed and true scores:

$$Q = \sum_{x=0}^{c-1} [\int_0^{\pi} P(x,p)dp] + \sum_{x=c}^{n} [\int_{\pi}^{1} P(x,p)dp], \tag{4}$$

where $P(x,p) = [\binom{n}{x} / B(\ell+1, m+1)]p^{\ell+x}(1-p)^{m+n-x}$ (Keats & Lord, 1962, p. 69).

The first integral in Equation 4 can be expressed as follows:

$$\int_0^{\pi} P(x,p)dp = [\binom{n}{x} / B(\ell+1, m+1)]\int_0^{\pi} p^{\ell+x}(1-p)^{m+n-x}dp$$

$$= [B(\ell+x+1, m+n-x+1)/B(\ell+x+1, m+n-x+1)][\binom{n}{x} / B(\ell+1, m+1)]\int_0^{\pi} p^{\ell+x}(1-p)^{m+n-x}dp$$

$$= [B(\ell+x+1, m+n-x+1)\binom{n}{x} / B(\ell+1, m+1)]\int_0^{\pi} [1/B(\ell+x+1, m+n-x+1)]p^{\ell+x}(1-p)^{m+n-x}dp$$

$$= [B(\ell+x+1, m+n-x+1)\binom{n}{x} / B(\ell+1, m+1)]I_{\pi}(\ell+x+1, m+n-x+1), \tag{5}$$

where $\binom{n}{x}$, $B(\ ,\ )$, and $I(\ ,\ )$ are respectively the binomial coefficient, the beta function, and the incomplete beta function.

Similarly the second integral in Equation 4 can be written:

$$\int_{\pi}^{1} P(x,p)dp = [\binom{n}{x} / B(\ell+1, m+1)]\int_{\pi}^{1} p^{\ell+x}(1-p)^{m+n-x}dp$$

$$= [B(\ell+x+1, m+n-x+1)\binom{n}{x} / B(\ell+1, m+1)]\int_{\pi}^{1} [1/B(\ell+x+1, m+n-x+1)]p^{\ell+x}(1-p)^{m+n-x}dp$$

$$= [B(\ell+x+1, m+n-x+1)\binom{n}{x} / B(\ell+1, m+1)][1-\int_0^{\pi} [1/B(\ell+x+1, m+n-x+1)]p^{\ell+x}(1-p)^{m+n-x}dp]$$

$$= [B(\ell+x+1, m+n-x+1)\binom{n}{x} / B(\ell+1, m+1)][1-I_{\pi}(\ell+x+1, m+n-x+1)]. \tag{6}$$

Finally, Equation 2 is obtained by substituting 5 and 6 into 4:

$$\underline{Q} = \sum_{\underline{x}=0}^{\underline{c}-1} [\underline{B}(\underline{\ell}+\underline{x}+1, \underline{m}+\underline{n}-\underline{x}+1)\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1, \underline{m}+1)] \underline{I}_{\underline{\pi}}(\underline{\ell}+\underline{x}+1, \underline{m}+\underline{n}-\underline{x}+1) +$$

$$\sum_{\underline{x}=\underline{c}}^{\underline{n}} [\underline{B}(\underline{\ell}+\underline{x}+1, \underline{m}+\underline{n}-\underline{x}+1)\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1, \underline{m}+1)] [1-\underline{I}_{\underline{\pi}}(\underline{\ell}+\underline{x}+1, \underline{m}+\underline{n}-\underline{x}+1)]$$

$$= \{1/\underline{B}(\underline{\ell}+1, \underline{m}+1)\}\{\sum_{\underline{x}=0}^{\underline{c}-1} \binom{\underline{n}}{\underline{x}} \underline{B}(\underline{\ell}+\underline{x}+1, \underline{m}+\underline{n}-\underline{x}+1) \underline{I}_{\underline{\pi}}(\underline{\ell}+\underline{x}+1, \underline{m}+\underline{n}-\underline{x}+1) +$$

$$\sum_{\underline{x}=\underline{c}}^{\underline{n}} \binom{\underline{n}}{\underline{x}} \underline{B}(\underline{\ell}+\underline{x}+1, \underline{m}+\underline{n}-\underline{x}+1) [1-\underline{I}_{\underline{\pi}}(\underline{\ell}+\underline{x}+1, \underline{m}+\underline{n}-\underline{x}+1)]\}.$$

References

Griffiths, D. A. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics, 1973, 29, 637-648.

Huynh, H. On consistency of decisions in criterion-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

LaValle, I. H. An Introduction to Probability, Decision, and Inference. New York: Holt, Rinehart and Winston, 1970.

Lord, F. M. A strong true score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

Pearson, K. (Ed.) Tables of the Incomplete Beta-Function. New Rochelle, N.Y.: Cambridge University Press, 1956.

Shenton, L. R. Maximum likelihood and the efficiency of the method of moments. Biometrika, 1950, 37, 111-116.

Subkoviak, M. J., & Albrecht, B. A. Empirical investigation of procedures for estimating reliability for mastery tests. Manuscript submitted for publication, 1977.

Wilcox, R. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics, 1977, in press.

Footnotes

Michael Subkoviak and Rand Wilcox were supported respectively by National Institute of Education Grants No. NIE-G-76-0088 and NIE-G-76-0083. Equal authorship is implied.

[1]Maximum likelihood estimates of $\underline{\ell}$ and $\underline{m}$ are also possible (see Griffiths, 1973). However, when the number of examinees is large, results reported by Shenton (1950) would seem to imply little difference between maximum likelihood estimates and the moment estimates of $\underline{\ell}$ and $\underline{m}$ given here.

Table 1

Values of $\binom{4}{x}$, $B(\ ,\ )$ and $I_{\pi}(\ ,\ )$ for the Numerical Example[a]

x	$\binom{4}{x}$	$B(9.76+x,13.76-x)$	$I_{.85}(9.76+x,13.76-x)$
0	1	$.12\times10^{-6}$	.99
1	4	$.95\times10^{-7}$	.99
2	6	$.87\times10^{-7}$	.99
3	4	$.95\times10^{-7}$	.99
4	1	$.12\times10^{-6}$	.99

[a] $B(9.76,9.76) = .15\times10^{-5}$

Further Comments on Reliability for Mastery Tests

Michael J. Subkoviak

University of Wisconsin

Running head: Classification Consistency for Mastery Tests

Abstract

This paper illustrates that the various coefficients of classification consistency that have been proposed as measures of reliability for mastery tests have different interpretations and statistical properties. As such, they should not be applied indiscriminately. Rather, a user should employ that coefficient that is most meaningful within the context of a particular problem.

Further Comments on Reliability for Mastery Tests

Huynh's recent critique can be reduced essentially to two points (1977). First, he argues that the kappa coefficient (κ) is preferable to the simple proportion of consistent classifications on two tests (P_C) as an index of reliability for mastery tests. Second, he notes that the procedure discussed by Huynh (1976) for estimating these indicies is mathematically more tractable than methods requiring the estimation of a testee's true ability (see Marshall & Haertel, 1976; Subkoviak, 1976).

Regarding the first point, it is not obvious that kappa is the only index of reliability that a user may wish to consider. The purpose of this paper is to demonstrate that coefficients like kappa or the proportion of consistent classifications on two tests have different interpretations and different statistical properties. Thus, such coefficients should not be used indiscriminately. Rather, a user should purposively select that coefficient whose properties are most suited to the intended application.

I quite agree with the second point regarding the utility of Huynh's procedure. In fact, since Huynh's method is based on a simple model for test scores proposed by Keats and Lord in 1962, similar procedures based on more recent and more sophisticated models are likely to provide even better estimates of reliability (see Lord, 1965, 1969).

Possible Reliability Coefficients for Mastery Tests

For simplicity, the following discussion is couched in terms of a mastery test, in which a single cutting score (C) is used to classify persons as masters or nonmasters. However, generalization to tests involving multiple cutting

scores and polychotomous classifications is immediate.

If a score of C or greater represents mastery on a test, then the proportion of persons consistently classified as master/master or nonmaster/nonmaster on two testings (P_C) is an indicator of the replicability of mastery-nonmastery outcomes. Goodman and Kruskal (1954, p. 758) proposed the use of P_C as a reliability coefficient; however, Huynh (1977) essentially rejects it on the grounds that P_C does not assume values between 1 and 0, as do traditional, norm-referenced reliability coefficients. Rather, P_C assumes values between 1 and P_{chance}, where $P_{chance} = [P^2(x \geq C) + P^2(x < C)] \geq 1/2$. $P(x \geq C)$ and $P(x < C)$ are the proportions of masters and nonmasters in the group tested.

Goodman and Krushal counter such objections as follows:

> Conventions like these [requiring that an index assume values between 1 and 0] have seemed important to some authors, but we believe they diminish in importance as the meaningfulness of the measure of association increases. One real danger connected with such conventions is that the investigator may carry over size preconceptions based upon experience with completely different measures subject to the same conventions. (p. 738)

Goodman and Kruskal conclude that the proportion of consistent classifications on two tests, P_C, is a meaningful and thus valid index of reliability. Furthermore, since the range of P_C is other than 1 to 0, norm-referenced standards of "good" and "poor" reliability are unlikely to be mistakenly applied to P_C.

For illustrative purposes, a unimodal distribution of scores (x) for a population on a four item test is shown in the first two columns of Table 1. Using the procedure discussed by Huynh (1976) and also by Keats and Lord (1962),

these single test administration data can be used to generate a bivariate distribution of scores (x and x') that would be expected for two test administrations (not shown). This bivariate scatterplot, in turn can be used to compute P_C, the proportion of consistent mastery/mastery and nonmastery/nonmastery outcomes on the two tests.

Insert Table 1 about here

For example, if the criterion of mastery is set at $C = 4$ in Table 1, the proportions of consistent mastery/mastery and nonmastery/nonmastery outcomes in the associated bivariate scatterplot are respectively $P(x \geq 4, x' \geq 4) = .01$ and $P(x < 4, x' < 4) = .85$, where x and x' represent scores on different test administrations. Thus, the total proportion of consistent outcomes when $C = 4$ is $P_4 = .01+.85 = .86$, as shown in the third column of Table 1. The proportions of consistent classifications for other values of C in Table 1 are similarly computed to be $P_3=.61$, $P_2=.61$, $P_1=.86$, and $P_0=1.00$.

Notice that the proportion of consistent classifications in the P_C column of Table 1 increases as the criterion (C) moves away from the central concentration of scores at $x=2$ into either tail of the distribution. In other words, P_C is a U-shaped function of C for this particular set of data.

Now suppose one wished to compare the observed proportion of consistent classifications on two tests, P_C, to the proportion of consistent outcomes expected if mastery-nonmastery decisions for each student were made instead by flipping a fair coin twice. Since the expected proportion of consistent decisions in the latter case is one-half (1/2), the difference $(P_C-1/2)$ indicates how much more consistent the actual decision process is than the random process just described. In addition, the transformation $(P_C-1/2)/(1-1/2) = 2P_C-1$ provides an index that assumes values between 1 and zero, if

desired. If the actual decision process is completely consistent then $2P_C$-1 equals 1; if the actual process is no better than the random process, then the index equals 0; generally the value is somewhere in between.

Values of $2P_C$-1 are shown in the fourth column of Table 1. Since $2P_C$-1 is a simple linear function of P_C, both coefficients are U-shaped functions of C for this particular data set. However, P_C is always greater than or equal to $2P_C$-1; so the same standards of "good" and "poor" cannot be applied to both.

Finally, one might again compare the observed proportion consistent outcomes, P_C, to the proportion expected by twice flipping a coin biased according to the relative proportions of masters and nonmasters in the group tested. If $P(x \geq C)$ and $P(x < C)$ are the proportions of masters and nonmasters in the population tested, then the expected proportion of consistent decisions in this case is $P_{chance} = [P^2(x \geq C) + P^2(x < C)] \geq 1/2$. Thus, the kappa coefficient, defined by $\kappa = (P_C - P_{chance})/(1-P_{chance})$, indicates how much more consistent the actual decision process is than this latter random process (Cohen, 1960). Kappa equals 1 if the actual process is perfectly consistent and equals zero if the actual process is no better than random (in the latter sense).

For example, if $C = 4$ in Table 1, then $P_{chance} = P^2(x \geq C) + P^2(x < C) = (.08)^2 + (.25+.34+.25+.08)^2 = .85$. Thus, $\kappa = (P_C - P_{chance})/(1-P_{chance}) = (.86-.85)/(1-.85) = .06$. The other values of kappa in Table 1 were similarly computed. Kappa is undefined at $C = 0$ because the term $(1-P_{chance})$ in the denominator of kappa equals zero in this case.

Notice in Table 1 that kappa as a function of C has an <u>inverted</u> U-shape--just the opposite of coefficients P_C and $2P_C$-1. Obviously then, consistency as measured by kappa is quite different from consistency as measured by P_C or

$2P_C$-1. But what accounts for this difference? The answer lies in the term P_{chance} that occurs in coefficient kappa.

P_C represents the total proportion of consistent mastery/mastery and nonmastery/nonmastery decisions that occur when a test is administered to a group composed of particular percentages of masters and nonmasters. If the group is largely composed of either masters or nonmasters, a major portion of the observed consistency, P_C, is attributable to the group constitution. (This is also true of $2P_C$-1 which is a linear function of P_C.) For example, if a group is largely made-up of masters, then consistent mastery/mastery decisions are quite likely to occur. Now the term $P_{chance} = P^2(x \geq C) + P^2(x < C)$ that occurs in coefficient kappa represents that portion of the observed consistency, P_C, that is due to the particular distribution of mastery and nonmastery in the group tested. Thus, coefficient kappa, $\kappa = (P_C - P_{chance})/(1-P_{chance})$, represents the proportion of consistent decisions attributable to factors other than group composition, such as the test itself, the conditions of administration, etc.

As such, the choice of P_C (or $2P_C$-1) vis-a-vis κ would seem to depend upon whether or not the user wishes to consider the group as an element of the decision process. In other words, if one is interested in the totality of consistent classifications that occur for a particular combination of test and group, then an index like P_C would seem to be appropriate. If one wishes to discount the effect of group composition on the decision process, then a kappa-like coefficient might be appropriate. It might be noted that there is ample precedent for an index, like P_C, that includes group effects. The conventional norm-referenced reliability coefficient, for example, is very much affected by the magnitude of true score variance for the particular group tested.

Sampling Error

The extent to which numerical estimates of an index vary from one sample to the next is another matter of some importance, particularly if such estimates are based on rather small, classroom size samples (N). For example, if random samples of 50 students were repeatedly drawn from the population of Table 1 and if estimates of the three coefficients ($\hat{P}_C$, $2\hat{P}_C-1$, and $\hat{\kappa}$) were repeatedly computed for each sample of test scores, the standard errors of the estimates would be approximately as shown in Table 2. For instance, if the criterion of mastery is $C = 4$, Table 2 indicates that estimates $\hat{P}_C$, $2\hat{P}_C-1$, and $\hat{\kappa}$ for samples of size 50 have associated standard errors of .05, .10, and .17 respectively.

Insert Table 2 about here

Comparing the values in Table 2, it is seen that the standard error of kappa tends to be generally larger than that of the other two estimators. This stems from the fact that kappa, $\kappa = (P_C - P_{chance})/(1-P_{chance})$, is a ratio of random variables; whereas the other two estimators are not. Thus, it would be important to insure adequate sample size, particularly when estimating kappa.

The values of Table 2 were obtained using the following approximations to the variance of the three estimators (Fleiss, Cohen, & Everitt, 1969; Hubert, 1977): (a) $\sigma^2(\hat{P}_C) = [P_C(1-P_C)]/N$, (b) $\sigma^2(2\hat{P}_C-1) = [4P_C(1-P_C)/N]$, and (c)

$$\sigma^2(\hat{\kappa}) = \{\sum_{i=1}^{2} P_{ii}[1-P_{chance}) - (P_{\cdot i}+P_{i\cdot})(1-P_C)]^2 + (1-P_C)^2 \sum_{\substack{i=1 \\ i\neq j}}^{2} \sum_{j=1}^{2} P_{ij}(P_{\cdot i}+P_{j\cdot})^2 - (P_C P_{chance} - 2P_{chance} + P_C)^2\}/[N(1-P_{chance})^4].$$

Here P_{ij} stands for the proportion in the ij^{th} cell of the 2 x 2 contingency table representing mastery-nonmastery outcomes on two testings; $P_{i\cdot}$ and $P_{\cdot j}$

represent the i^{th} row and j^{th} column (marginal) proportions of the 2 x 2 table; and the other symbols are as previously defined. It should be noted that these formulae are based on the assumption that a binomial distribution is responsible for generating the 2 x 2 contingency table and that the sample of students is reasonably large (see Hubert, 1977).

Conclusion

The basic message of this paper is that the various measures of classification consistency thus far proposed for mastery tests have different interpretations and statistical properties. As such, they should not be used blindly. Rather, a user should deliberately select that coefficient that is most meaningful within the context of a particular problem. In making this decision, the user might also wish to consider coefficients that reflect nearly consistent classifications (Goodman & Kruskal, 1954, p. 758) or coefficients that reflect the stability over repeated testing of score deviations about criterion C (Brennan & Kane, 1977).

References

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, in press.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 1969, 72, 323-327.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. Journal of the American Statistical Association, 1954, 49, 732-764.

Hubert, L. J. Kappa Revisited. Psychological Bulletin, 1977, in press.

Huynh, H. Reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Huynh, H. Reliability of criterion-referenced tests: Comments on a paper by Subkoviak. Journal of Educational Measurement, 1977, submitted.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

Marshall, J. L., & Haertel, E. H. The mean split-half coefficient of agreement: A single administration index of reliability for mastery test. Unpublished manuscript, 1976. (Available from Department of Educational Psychology, 1025 West Johnson St., Madison, Wisconsin 53706).

Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 265-276.

Table 1

Reliability as a Function of

Criterion Score for Various Coefficients

x or C	$f(x)$	P_C	$2P_C-1$	κ
0	.08	1.00	1.00	undefined
1	.25	.86	.72	.06
2	.34	.61	.22	.11
3	.25	.61	.22	.11
4	.08	.86	.72	.06

Table 2

Estimates of Standard Error for Samples of 50 Students

$\underline{C}$	$\sigma(\hat{P}_{\underline{C}})$	$\sigma(2\hat{P}_{\underline{C}}-1)$	$\sigma(\hat{\kappa})$
1	.05	.10	.17
2	.07	.14	.14
3	.07	.14	.14
4	.05	.10	.17

Confirmatory Inference and Geometric Models

Lawrence J. Hubert

Michael J. Subkoviak

The University of Wisconsin, Madison

Confirmatory Inference and Geometric Models

Abstract

A confirmatory method is discussed for comparing an outside variable to a given geometric model, or alternatively, to the raw data from which the model is derived. The inference procedure is based on relatively simple nonparametric principles and requires the comparison of a proximity matrix generated from a geometric representation against a second "structure" matrix obtained from the outside variable under study. A number of examples are presented that illustrate how the same statistical approach can be applied in evaluating geometric models that arise in a number of ways, for instance, those produced by some explicit data reduction process, or possibly, models generated by naturally occuring spatial contiguity.

Confirmatory Inference and Geometric Models

Introduction

In the literature on data analysis over the last twenty years, a distinction between exploratory and confirmatory procedures has become very popular (see Kaiser, 1970). Supposedly, an exploratory strategy involves the use of some analysis technique on a given data set with the aim of identifying interesting relationships, patterns, and the like. On the other hand, a confirmatory approach requires the statement of a rather strong a priori conjecture which is then tested directly against the available data. It is assumed that these latter hypotheses are derived from a source outside the data actually used for the purposes of validation, possibly from the results of some previous exploratory study. Unfortunately, the word "exploratory" has gained such legitimacy that it now serves too often as a way of justifying isolated empirical studies that an investigator has no intention whatsoever of pursuing further, or more seriously, as a cover for a superficial theoretical conceptualization and a hurried research agenda.

Given the influence of classical statistics on the conduct of experimental studies in the behavioral sciences, the trend toward a confirmatory approach to research was very strong after World War II. More recently, however, inexpensive exploratory computational routines have become widely available, which has encouraged some attitude of "fadishness" in the analysis of data, irrespective of whether the methods chosen are appropriate for the problem or could possibly lead to any increased understanding of the area under study. In particular,

methodology to the novice can easily seem more important than substance when faced with the vast array of very elegant data reduction techniques, such as multidimensional scaling, cluster analysis, and similar paradigms. Too often the intricacies of a methodology become more crucial than the original research question, and in fact, even for substantively oriented investigators, there is some danger of using data more as a vehicle for displaying a specific statistical method than as the major reason for which an empirical analysis is carried out in the first place.

Although it may be obvious that confirmatory analyses would be desirable as an adjunct to many of the current approaches used in the study of proximity matrices (such as clustering and multidimensional scaling), very few techniques have been proposed in the literature that could help carry out such a program with any degree of rigor. Typically, users of the newer data reduction procedures operate more or less atheoretically, or at best, try to interpret their results in terms of intuitively reasonable arguments based on outside information regarding the objects or entities being studied. It is somewhat surprising that this same information which can be invoked in explaining the results of an analysis in a post-hoc fashion is not being used more directly in some confirmatory manner and without the imposition of an intermediate data reduction process.

With the motivation in mind of using more fully whatever outside information is available for a given data set, this paper will attempt to review a number of approaches to the analysis of proximity data based on confirmatory principles. Although it is accepted that exploratory strategies are of interest in generating insight and formal hypotheses for later verification, it would be of value at the same time to have a more complete set of confirmatory procedures that could be used in evaluating or supplementing exploratory results

in a rigorous fashion. The discussion to follow is organized into several major subsections that illustrate how specific data analysis problems can be attacked within a common nonparametric confirmatory strategy. In particular, to limit the scope of the discussion, our emphasis will be on geometric models, or more specifically, on data representations that have some explicit geometric interpretation. Within this context, our aim is to demonstrate how an obtained geometric representation may be evaluated against available outside information, or alternatively, how such representations in some cases might be bypassed altogether. Since parts of this material are not entirely new but are scattered throughout the literature, appropriate references will be included for the reader interested in pursuing a more complete presentation of the various topics introduced.

Confirmatory Strategies and Geometric Models

As a tactic for explaining how a confirmatory approach to data analysis might be carried out, a number of specific problems will be discussed illustrating the necessary concepts in a concrete manner. In particular, four topic areas are introduced in the sections to follow that demonstrate how a specific confirmatory strategy based on very simple nonparametric principals may be applied in several different ways. Depending on the context, it is conceivable that our method might be used instead of a geometric model; in extending an existing analysis strategy based on geometric notions; as a means of interpreting a given model with respect to outside information; or finally, as a preliminary to the construction of a desired geometric representation. The first example below formalizes the basic ideas to be used throughout the paper, and specifically, illustrates how a geometric model might be bypassed altogether when strong a priori conjectures are available.

Example 1

In a recent study concerned with the "goodness" of patterns, Glushko (1975) attempted to verify Garner's (1962) basic hypothesis regarding what makes one pattern better than another. To be more specific, each of the 17 patterns used by Glushko, listed in Table 1, can be characterized by the size of an inferred equivalence class. The term "equivalence" is used to label the set of patterns that contain a single figure plus all other configurations that result from reflections and/or 90 degree rigid rotations. As indicated in Table 1, two of the Glushko patterns construct the same configuration under all of these operations, 8 patterns have 4 associated figures, and finally, 7 patterns produce 8 different members in its class. According to Garner, pattern goodness is a direct function of the size of a configuration's inferred equivalence class, with the smaller size classes corresponding to the better patterns.

Insert Table 1 about here

To test Garner's hypothesis using the 17 figures of Table 1, Glushko, first of all, obtained a symmetric measure of proximity between each pair of patterns using a choice task. Twenty subjects were presented all 136 different pattern combinations and were asked to indicate preference. These choices were then summed over subjects and subtracted from an expected preference frequency of 10. The absolute values of these differences, given in the lower triangular portion of Table 2, form a symmetric measure of proximity defined for all pattern pairs and provide data in a form that can be subjected to a variety of

---

Insert Table 2 about here

---

data reduction techniques. In particular, Glushko attempted to represent the

structure of the proximity function by, first of all, placing the 17 configurations in a two-dimensional space using Kruskal's (1964a,b) well-known multidimensional scaling routine. Given this geometric representation, Johnson's (1967) diameter clustering results were then superimposed, producing a representation similar to that we give in Figure 1 (here, we only indicate the clustering result defined by three subclasses). Clearly, one strong dimension (the vertical) can be identified as that of equivalence class size. In addition, the clusters themselves correspond fairly well to a grouping on the basis of the same criterion except for the minor misplacement of the two configurations numbered 10 and 11.

Insert Figure 1 about here

The process of verifying Garner's hypothesis through a multidimensional scaling and clustering seems rather circuituous, especially since the equivalence class hypothesis implies a definite structure for the original proximity measure. Although the clustering and scaling results in this case are clear-cut, unambiguous outcomes of this type are rare, to say the least. Unfortunately, when a strong hypothesis is not reflected as dramatically in the scaling or clustering results, it may be difficult to decide whether the hypothesis is inadequate or the data reduction techniques are at fault. In the typical application, the researcher may be able to identify portions of his theory in a scaling or clustering solution but lacks a strategy for measuring in any precise manner the actual degree of confirmation or nonconfirmation.

As an alternative approach, it should be possible to test in a direct manner whether the pattern goodness hypothesis is reflected in the original proximities and bypass the scaling and clustering solutions altogether. To introduce some notation, suppose we denote the patterns as $o_1, o_2, \ldots, o_n$ (where

n is 17 in our example). Furthermore, let $q(o_i,o_j)$ refer to the symmetric proximity between patterns o_i and o_j, and $\underset{\sim}{Q}$ to an organization of these measures into a 17 by 17 square matrix with rows and columns labeled by the objects or patterns $o_1, o_2, \ldots, o_n$. By convention, the diagonal of $\underset{\sim}{Q}$ is assumed to consist entirely of zeros. In addition to the empirical proximity matrix $\underset{\sim}{Q}$, the stated hypothesis will be represented numerically by a second "structure" matrix $\underset{\sim}{C}$ with elements $c(o_i,o_j)$. Explicitly, suppose $N(o_i)$ denotes the size of the inferred equivalence class for object o_i, and let f be some monotone function on the integers, e.g., $f(x) > f(y)$ if and only if $x > y$. Then, as a formal definition,

$$c(o_i,o_j) = f(|N(o_i) - N(o_j)|),$$

where it is assumed that $c(o_i,o_j) = 0$ for $o_i = o_j$. Although many functions f could be used and the actual choice will depend on the researcher's judgment as to the most appropriate relative size of the structure values, for the purposes of an illustration, f is taken as the identity, i.e., $f(x) = x$. In other words, the symmetric function values $c(o_i,o_j)$, given in the upper triangular portion of Table 2, are merely the absolute values of the differences in equivalence class sizes associated with the objects o_i and o_j.

As an operational interpretation, the theory used to generate the function $c(o_i,o_j)$ is given empirical support if the two sets of elements $c(o_i,o_j)$ and $q(o_i,o_j)$ have a similar patterning of high and low entries. Although many formal indices for this relationship could be defined, the pairing of a proximity $q(o_i,o_j)$ with a structure value $c(o_i,o_j)$ suggests that the simple Pearson product-moment correlation may be a natural measure to consider; and thus, will be our choice for the sequel. Once this index is calculated, the next problem concerns its significance, and specifically, with whether the size of the observed correlation between the values for $q(o_i,o_j)$ and $c(o_i,o_j)$ is sufficient

to reject some appropriately defined null hypothesis.

To generate a reasonable reference distribution for the observed correlation, suppose a randomness hypothesis is assumed that hopefully can be rejected. More specifically, it is conjectured that the partition of the objects (or patterns) $o_1, o_2, \ldots, o_n$ occurred randomly or was chosen at random from the set of all partitions of the same form. In our case, the conjectured partition contains three classes with 2, 8, and 7 objects in each, and thus, the null hypothesis of interest asserts that this particular partition occurred randomly, and consequently, does not reflect the patterning of entries in the proximity matrix $\underset{\sim}{Q}$. Moreover, any such partition of $\underset{\sim}{Q}$ of the same form (i.e., number of classes) will produce a correlation index and when completely enumerated will generate an exact reference distribution for the assumed "null" hypothesis. From an inference perspective, the observed correlation for the conjectured partition can be compared to this distribution and if at a suitably extreme percentage point, the "null" hypothesis of randomness can be rejected. In short, whenever the correlation actually obtained for the conjectured partition is large enough, then this index can be assumed to reflect a value that was obtained nonrandomly, i.e., at least to some extent, the functions $q(o_i, o_j)$ and $c(o_i, o_j)$ have a common patterning of high and low entries.

Although complete enumeration is usually prohibitive because of computational costs, and thus, an exact reference distribution is typically too expensive to obtain, Monte Carlo approximations are relatively inexpensive (cf. Hubert and Schulz, 1975; Schultz and Hubert, 1976). For instance, Table 3 presents the frequency results of selecting 1000 partitions of the desired form at random and with replacement, and should provide an approximate distribution that is fairly accurate for this application. In particular, using the Table 1

Insert Table 3 about here

data, the observed correlation for the Garner hypothesis is .640, which is greater than any value observed in the Table 3 distribution. Thus, the null hypothesis of a random partition can be rejected at an approximate significance level of .000, suggesting that the equivalence class hypothesis is supported by the patterning of the proximity values.

Using the previous example as a guide, the salient features of a confirmatory analysis should be evident. Given a proximity measure $q(o_i,o_j)$ and some conjecture specified in terms of a structure function $c(o_i,o_j)$, the observed correlation between $q(o_i,o_j)$ and $c(o_i,o_j)$ is compared to a reference distribution generated under a hypothesis of randomness. If the obtained correlation is at an extreme percentage point, the correspondence between $q(o_i,o_j)$ and $c(o_i,o_j)$ is declared "significant", with the added implication that the conjecture leading to the construction of $c(o_i,o_j)$ may help "explain" some of the variation present in the empirical proximity measures.

Although the example given above implies that a randomness hypothesis should be defined in terms of selecting a partition of a given form at random, a more general hypothesis can also be considered that will generate exactly the same distribution. Explicitly, if the values assigned by the proximity function are organized, as before, into an n x n square matrix $\underset{\sim}{Q}$, and similarly, the values of the structure function into a second n x n square matrix $\underset{\sim}{C}$, both with rows and columns labeled as $o_1,o_2,\ldots,o_n$, then each reordering of the rows and simultaneously the corresponding columns of $\underset{\sim}{Q}$ in relation to the fixed $\underset{\sim}{C}$ matrix will induce a specific partition of the n objects $o_1,\ldots,o_n$. In other words, for our $\underset{\sim}{C}$ matrix of Table 2, any reordering of $\underset{\sim}{Q}$ produces a partition

defined by subsets containing 2, 8 and 7 objects. The first two rows and columns of the reordered $\underset{\sim}{Q}$ matrix define the objects in the class of size 2, the next 8 rows and columns define a class of 8 objects, and the remaining 7 rows and columns define the last object class of size 7. Moreover, if a reordering of $\underset{\sim}{Q}$ is chosen at random, that is all n! possible reorderings are considered equally likely, then this assumption induces a random selection of a partition of the same general form used in the original construction of the $\underset{\sim}{C}$ matrix. In short, the random reordering of $\underset{\sim}{Q}$ and the random selection of a partition will generate exactly the same distribution of correlations, and thus, either concept may be used in producing an approximate reference table through Monte Carlo simulation. This generalization will prove very important later when a confirmatory approach is necessary but one that cannot be identified by a specific partitioning of an object set.

Although we suggest carrying out a confirmatory test through the use of an approximate distribution obtained through Monte Carlo simulation, it is also possible to find the <u>exact</u> mean and variance of the complete reference distribution by formula, given only the matrices $\underset{\sim}{C}$ and $\underset{\sim}{Q}$. Specifically, the mean of the Pearson correlation r is 0 and its variance equal to

$$V(r) = \{1/\binom{n}{2}\}\{1 + \frac{1}{(n-2)G_2H_2}[2(G_1-G_2)(H_1-H_2)$$

$$+ \frac{(2G_1-G_2)(2H_1-H_2)}{(n-3)}]\},$$

where

$$G_1 = \sum_{i}\sum_{i\neq j}(q(o_i,o_j) - \bar{q})^2;\quad G_2 = \sum_{\substack{i\\ i\neq j}}\sum_{j}(q(o_i,o_j) - \bar{q})^2;$$

$$H_1 = \sum_{i}\sum_{i\neq j}(c(o_i,o_j) - \bar{c})^2;\quad H_2 = \sum_{\substack{i\\ i\neq j}}\sum_{j}(c(o_i,o_j) - \bar{c})^2,$$

and

$$\bar{q} = \frac{2}{n(n-1)} \sum_{i<j} \sum q(o_i,o_j); \quad \bar{c} = \frac{2}{n(n-1)} \sum_{i<j} \sum c(o_i,o_j).$$

As an example of how the variance calculation may be used for the data of Table 1 and the structure function of Table 2, we find $\sqrt{V(r)} = .0879$. Converting to a Z-score for the observed correlation of .640, a value of 7.28 is obtained, which would indicate a rather significant result if it were possible to assume even a crude normality (see Mantel, 1967, for the appropriate moment derivations).

Example 2

The previous example illustrates how a confirmatory approach might be used in directly verifying a stated a priori conjecture against the original set of proximities. As an alternative application of these same principles, it is also possible to extend several analysis strategies proposed in the literature that are based on naturally occurring geometric models. Most of the appropriate references relate to the organization of a group of objects, e.g., people, census tracts, and so on, derived from some notion of geographic or spatial contiguity. Within this context, one of the major analysis tasks concerns the association between spatial contiguity and some other variable or variables measured on these same objects.

As a concrete example from social psychology, Campbell, Kruskal, and Wallace (1966) developed an index of seating aggregation and an associated significance testing strategy for determining whether the observed black-white seating adjacencies within a classroom might be considered random. The geometric model in this case is defined by the occupied seats within a classroom, or more specifically, by the spatial location of the students in a two-dimensional plane. The outside variable of interest is dichotomous, i.e., black or white, and the inference task is one of determining whether the spatial positioning of blacks and whites indicates aggregation, e.g., whether blacks sit

with blacks and whites sit with whites.

The confirmatory approach developed in the previous section includes the Campbell, et al. approach as a special case. In particular, the set of objects $\{o_1,o_2,\ldots,o_n\}$ now refers to the set of n people and $q(o_i,o_j)$ refers to some measure of spatial distance obtained from the observed seating pattern. Although very general measures of distance could be used, Campbell, et al. consider a simple index defined (in our notation) as follows:

$$(1)\quad q(o_i,o_j) = \begin{cases} 1 & \text{if } o_i \text{ and } o_j \text{ are seated adjacently within a single row;} \\ 0 & \text{otherwise.} \end{cases}$$

For the Campbell, et al. application, the structure function $c(o_i,o_j)$ would be obtained from the outside variable of race:

$$(2)\quad c(o_i,o_j) = \begin{cases} 1 & \text{if } o_i \text{ and } o_j \text{ are both black or both white;} \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, the cross-product statistic $\sum_{i<j} q(o_i,o_j)c(o_i,o_j)$, which forms the crux of the correlation coefficient between the functions $c(o_i,o_j)$ and $q(o_i,o_j)$, is the number of same-race adjacencies observed in the given seating pattern. Using this statistic, evidence for aggregation is indicated by a large value, or in our correlational context, by a large positive correlation between $q(o_i,o_j)$ and $c(o_i,o_j)$.

Although the $c(o_i,o_j)$ function used above is rather simple, it can be generalized to include variables that represent more than a simple dichotomy. For example, suppose $x_1,x_2,\ldots,x_n$ denote some numerical variable attached to each of the n objects and define

(3) $c(o_i,o_j) = \max\{x_1,\ldots,x_n\} - |x_i - x_j|$.

If, for instance, $x_k = 0$ when o_k is black and 1 when o_k is white, then $\max\{x_1,\ldots,x_n\} = 1$, and this function is equivalent to that given in (2). More significantly, if x_k denotes, say, the age of person o_k, then this same type of function could be used to relate the information contained in this variable to seating pattern. In other words, when x_k refers to age, then large positive correlations between $q(o_i,o_j)$ defined as in (1) and $c(o_i,o_j)$ defined as in (3) will suggest that people of a similar age sit together. Obviously, many other functions of the variables $x_1,\ldots,x_n$ could be used for defining $c(o_i,o_j)$, and also, other notions of spatial distance could be used in defining $q(o_i,o_j)$. What is most important, however, is the general appropriateness of the Monte Carlo significance testing strategy and the variance term discussed in the previous section.

The concepts presented in this example can be extended to several other situations of interest in social psychology that have been developed in slightly different directions. For instance, suppose the n objects are now census tracts and the task is to relate an outside variable to contiguity of the tracts defined geographically (see Cliff & Ord, 1971; Geary, 1954; Royaltey, Astrachan, & Sokal, 1975). The same analysis procedure is appropriate with geographical distance defining $q(o_i,o_j)$ and the function $c(o_i,o_j)$ defined, say, as in (3), or possibly, by some other more complicated function of a vector of variables available for each object. Also, the analysis strategy for observations connected through some general network structure, e.g., kinship, can be approached

in the same way and related to outside data, e.g., to socioeconomic status. In fact, this topic as defined by Winsborough, Quarantelli, and Yutzy (1963) is also a special case of our confirmatory analysis strategy.

Example 3

Geometric configurations that are generated as a result of a multidimensional scaling (see Carroll & Chang, 1970; Kruskal, 1964a,b) represent yet another context in which the confirmatory paradigm could be used to test a priori conjectures. Even though these spatial representations are produced by an explicit data reduction process and consequently do not arise naturally as in Example 2, some of the same hypothesis testing principles are still appropriate. To give an illustration, consider the application of Carroll and Chang's individual differences scaling procedure to data collected by Wish, Deutsch, and Biener (1970). The objects of study for this analysis were 12 nations; and since each of 18 subjects rated the proximity of all pairs of nations on a nine point scale (large numbers indicating a greater degree of similarity), the resulting 18 proximity matrices, all of size 12 x 12, are appropriately analyzed by the Carroll-Chang routine (1970). The group result selected for our discussion is a two dimensional configuration, shown in Figure 2, in which each nation is represented by a point, and where the inter-

---

Insert Figure 2 about here

---

point distances reflect the degree of similarity between the corresponding nations as judged by the group, e.g., the distance between the U.S. and China is large since they are perceived on the average as being very dissimilar.

Instead of attempting to label and interpret dimensions per se, suppose the researcher wishes to test the a priori hypothesis that an outside variable,

such as political alignment, accounts in part for the distances between nations. In other words, the researcher is interested in confirming the conjecture that nations close together subscribe to similar political philosophies, and conversely, those far apart have different political systems. In this case the proximity function $q(o_i,o_j)$ would merely refer to the distance between nations o_i and o_j in Figure 2. Or more specifically, if (y_{i1},y_{i2}) denotes the numerical coordinates computed by the Carroll-Chang procedure for the point o_i in Figure 2, then the Euclidean distance between any two points o_i and o_j in the figure is defined as

$$(4) \quad q(o_i,o_j) = [(y_{i1}-y_{j1})^2 + (y_{i2}-y_{j2})^2]^{1/2}.$$

As in previous examples, the structure function $c(o_i,o_j)$ would be obtained from the outside variable of political alignment. For instance, if political alignment were simply dichotomized as communist vs. noncommunist, then $c(o_i,o_j)$ might be defined as

$$(5) \quad c(o_i,o_j) = \begin{cases} 0 & \text{if } o_i \text{ and } o_j \text{ are both communistic or both noncommunistic;} \\ 1 & \text{otherwise.} \end{cases}$$

With this notation, a large positive correlation between $q(o_i,o_j)$ as defined in (4) and $c(o_i,o_j)$ as defined in (5) would indicate that nations of similar political persuasion are located close together in Figure 2. As it turns out, the observed correlation between the interpoint distances in Figure 2 and the dichotomous variable of political alignment given by (5) is .50, which is significant at the .000 level (approximately) when referred to the distribution of correlations for 1000 random reorderings of matrix $\underset{\sim}{Q}$. In short, there is convincing statistical support for the hypothesis that political alignment partially accounts for the arrangement of points. This conclusion is consistent with the descriptive analysis conducted by Wish et al. (1970), which identified

the vertical dimension of Figure 2 as a political alignment factor with communist nations at the top. In a similar manner an outside variable such as gross national product could be used to support the Wish et al. contention that the horizontal axis of Figure 2 might be identified as "economic development", e.g., the more highly developed nations are located to the right. Again, however, our aim would be to relate the gross national product index directly to the interpoint distances, and thus, any explicit dimensional representation would be ignored.

In the example given above, political alignment was related to the interpoint distances of Figure 2 as derived from a particular data reduction procedure. However, since the distance between two points in Figure 2 is simply a graphic representation of the rated similarity between two nations, the political alignment hypothesis could also be tested directly against subjects' raw proximity ratings, thus bypassing Figure 2 and the Carroll-Chang analysis altogether. For instance, the mean similarity ratings for each pair of nations in the Wish et al. study are as shown in Table 4, with larger values now indicating greater similarity. Thus, if Table 4 is used to define a new

---

Insert Table 4 about here

---

proximity function $q(o_i,o_j)$ and if (5) again represents the structure function $c(o_i,o_j)$, then a large negative correlation between $q(o_i,o_j)$ and $c(o_i,o_j)$ would directly indicate that political alignment plays an important part in the formation of subjects' similarity judgments, i.e., nations which have the same political systems also receive high similarity ratings. The actual correlation between $q(o_i,o_j)$ and $c(o_i,o_j)$ for these data is -.34, which is significant at an approximate .013 level when referred to the distribution of correlations for 1000 random permutations of Table 4. In summary, the political alignment

hypothesis can be tested either against Figure 2 or against the raw proximity data, with the latter being somewhat simpler and more direct if the researcher is willing to sacrifice the advantage of a pictorial representation.

In addition to the geometric configuration of nations given in Figure 2, the Carroll-Chang procedure also produces a configuration of the particular subjects that supplied the similarity data, as shown in Figure 3. The horizontal and vertical axes of Figure 3 are exactly the same as those of Figure 2, representing political alignment and economic development, respectively. Numerical

---

Insert Figure 3 about here

---

coordinates (w_{i1},w_{i2}) again locate a subject o_i in Figure 3, and furthermore, indicate how much emphasis a subject o_i gives to political alignment and economic development when rating the similarities of nations. For instance, referring to Figure 3, subject 10 gives primary emphasis to the economic development dimension, subject 11 gives primary emphasis to political alignment, and subjects in the center of the configuration weight both dimensions about equally.

As indicated in Figure 3, Wish et al. further classified each subject either as a hawk (H), moderate (M), or a dove (D) according the person's stance on the Vietnam War, and descriptively argue that subjects in the same class tend to weight the two dimensions similarly. In other words, since it is hypothesized that hawks, moderates, and doves will form reasonably homogeneous clusters in Figure 3, the confirmatory paradigm provides a statistical test for the conjecture that subjects weight dimensions differentially according to their political opinions. Again, the proximity function is defined as the Euclidean distance between points o_i and o_j in Figure 3:

$$(6) \quad q(o_i,o_j) = [(w_{i1}-w_{j1})^2 + (w_{i2}-w_{j2})^2]^{1/2}$$

and for the sake of simplicity, the structure function is defined as

$$(7)\quad c(o_i,o_j) = \begin{cases} 0 & \text{if } o_i \text{ and } o_j \text{ belong to the same class (hawk, moderate, or dove);} \\ 1 & \text{otherwise.} \end{cases}$$

A large positive correlation between the function values $q(o_i,o_j)$ and $c(o_i,o_j)$ given in (6) and (7) supports the conjecture that hawks, moderates, and doves tend to form separate clusters in Figure 3. Since the observed correlation is .19, which is significant at an approximate .009 level, the hypothesis is given statistical support. Wish et al. note specifically that hawks tend to cluster above the diagonal in Figure 3 and give relatively more emphasis to the political alignment factor; whereas moderates and doves cluster below the diagonal and give relatively more weight to economic development.

Although it should be clear that exploratory analyses such as multidimensional scaling may generate a rich source of hypotheses, and the confirmatory paradigm may provide a useful means for subsequently testing such hypotheses more formally, a word of explicit caution is also in order. Specifically, a hypothesis arising from an exploratory analysis of a particular data set should not be tested on the same set of data, since such a strategy amounts to "data snooping" and may produce significant results that cannot be replicated. Typically, if the researcher is bound by a single data set it may still be possible to take advantage of both exploratory and confirmatory analyses by randomly dividing the data in half. Exploratory analyses could then be applied to one half of the data, generating interesting research questions, which could be tested on the second half using the confirmatory paradigm. For instance, the 18 subjects in the Wish et al. study could be assigned in equal numbers to two groups, and the Carroll-Chang procedure applied to the data of one group, giving rise to a representation such as Figure 2 and the associated political

alignment conjecture. The confirmatory paradigm could then be applied to the remaining data, leading to either acceptance or rejection of the a priori hypothesis.

Example 4

The application of the confirmatory paradigm illustrated in Example 3 can easily be generalized beyond the specific context of multidimensional scaling and used to investigate data that is more traditionally considered within an analysis of variance design. As an illustration, suppose $(y_{i1},y_{i2},\ldots,y_{ir})$ represents a profile of r different measurements on an object o_i. As in the previous example, o_i can be represented as a point in r-dimensional space defined by these numerical coordinates, and furthermore, if the r measures are commensurate or have been made so by an appropriate standardized transformation, then a Euclidean (or other) distance between objects o_i and o_j in r-dimensional space could be used to define a proximity function of the form

$$(8)\quad q(o_i,o_j) = \left[\sum_{k=1}^{n} (y_{ik}-y_{jk})^2\right]^{1/2}.$$

As a numerical example, Table 5 contains simulated profiles for n = 21 objects (persons) on r = 3 variables (e.g., standardized tests) taken from Mielke, Berry, & Johnson (1976, p. 1419). The three sets of measures are commensurate in the sense that all have the same range, and thus, may be substituted directly

---

Insert Table 5 about here

---

into (8) to obtain the distance between any object pair o_i and o_j.

The objects of Table 5 have been partitioned into four distinct subgroups on the basis of some outside variable, e.g., freshman/sophomore/junior/senior, and as in the previous examples, this outside variable can be used to define a structure function such as:

$$(9)\quad c(o_i,o_j) = \begin{cases} 0 & \text{if } o_i \text{ and } o_j \text{ are members of the same subgroup;} \\ 1 & \text{otherwise.} \end{cases}$$

If the confirmatory paradigm is used to test the hypothesis that the outside variable accounts in part for the arrangement of points in r-dimensional space, then a large positive correlation between $c(o_i,o_j)$ as defined in (9) and $q(o_i,o_j)$ as defined in (8) would support the conjecture that students in the same academic year tend to have similar test profiles, i.e., they tend to be close together in three dimensional space. In the case of Table 5, this correlation is .55, which is significant at an approximate .000 level. In other words, the comparison of $q(o_i,o_j)$ and $c(o_i,o_j)$ essentially carries out a multivariate analysis of variance involving four groups and r measurements on each subject in a group. Furthermore, even though the outside variable in this example is treated as a simple categorical measure, it should be clear from the discussion related to the structure function in (3) that an analogous function could be defined for variables measured on higher order scales. In fact, we could explicitly take the freshman/sophomore/junior/senior ordering into account in our illustration.

Discussion

As should be evident in the examples given above, the confirmatory approach developed in this paper has a number of applications related to the use and development of geometric models, either those that occur naturally or those derived from some intermediate data reduction process. In addition to the illustrations provided, a number of other correspondences to the methodological literature of the behavioral sciences could be developed that the reader may be

interested in pursuing further. For instance, Carroll and Chang (Note 1) suggest a general index of nonlinear correlation between two sets of observations $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_n\}$ defined by

$$K = (1/S_Y^2) \sum_{i \neq j} \sum w_{ij}(y_i-y_j)^2,$$

where

$$S_Y^2 = (1/n) \sum_{i=1}^{n} (y_i-\bar{y}.)^2,$$

and

w_{ij} is some decreasing monotonic function of $|x_i-x_j|$.

Intuitively, the smaller K is, the greater the inferred nonlinear correlation. Since S_Y^2 is constant over all permutations of the y's, a permutation distribution for K can be obtained by considering only its numerator, treating the values w_{ij} as a $\underset{\sim}{Q}$ matrix, and $(y_i-y_j)^2$ as a $\underset{\sim}{C}$ matrix. As discussed by Carroll and Chang, K itself includes, as special cases, the well-known correlation ratio as well as von Neuman's autocorrelation statistic. In fact, as a second general application, a substantial literature in sociology exists in using what is called the contiguity ratio, which is almost identical in form to K, as a way of relating the structure of social networks to outside variables (e.g., see Winsborough, Quarantelli, & Yutzy, 1963; Althauser, Burdick, & Winsborough, 1966). For a number of other applications of the type of analysis discussed in this paper but which are not specifically tied to geometric modeling, the reader should consult Schultz and Hubert (1976), Hubert and Baker (1977), Hubert and Levin (1976a,b), Hubert and Schultz (1976), and Hubert (1977).

Table 1

The Patterns Used by Glushko in Testing Garner's Pattern Goodness Hypothesis

Equivalence class size		
1	4	8
(1)	(3)	(11)
(2)	(4)	(12)
	(5)	(13)
	(6)	(14)
	(7)	(15)
	(8)	(16)
	(9)	(17)
	(10)	

Table 2

The Symmetric Proximity Matrix Obtained by Glushko for the Patterns of Table 1 (Lower Triangle) and the Structure Matrix Generated by the Equivalence Class Hypothesis (Upper Triangle)

Pattern	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	X	0	3	3	3	3	3	3	3	3	7	7	7	7	7	7	7
2	1	X	3	3	3	3	3	3	3	3	7	7	7	7	7	7	7
3	1	2	X	0	0	0	0	0	0	0	4	4	4	4	4	4	4
4	2	4	0	X	0	0	0	0	0	0	4	4	4	4	4	4	4
5	3	3	1	1	X	0	0	0	0	0	4	4	4	4	4	4	4
6	2	4	1	1	1	X	0	0	0	0	4	4	4	4	4	4	4
7	2	4	3	2	1	2	X	0	0	0	4	4	4	4	4	4	4
8	3	5	2	1	2	1	0	X	0	0	4	4	4	4	4	4	4
9	4	4	2	1	5	3	3	4	X	0	4	4	4	4	4	4	4
10	4	5	4	4	3	3	3	5	4	X	4	4	4	4	4	4	4
11	5	5	3	4	3	0	2	3	1	1	X	0	0	0	0	0	0
12	5	6	4	6	4	1	5	5	2	1	3	X	0	0	0	0	0
13	6	7	7	6	5	4	5	6	5	1	4	1	X	0	0	0	0
14	7	6	4	4	5	4	6	5	4	2	4	1	1	X	0	0	0
15	6	7	5	7	4	5	5	4	5	0	3	0	0	1	X	0	0
16	7	8	5	5	6	4	4	3	4	1	4	2	2	0	1	X	0
17	7	7	5	5	5	6	5	4	6	3	6	2	3	1	1	1	X

Table 3

Approximate Distribution for the Comparison of the Structure and Proximity Matrices Given in Table 2 (Sample Size of 1000)

Correlation	Sample Cumulative Proportion
-.193	.001
-.171	.005
-.162	.010
-.117	.050
-.098	.100
-.070	.200
-.046	.300
-.025	.400
-.009	.500
.010	.600
.033	.700
.068	.800
.115	.900
.162	.950
.273	.990
.297	.995
.396	.999
.420	1.000

Table 4

Mean Similarity Ratings for 12 Nations[a]

	1	2	3	4	5	6	7	8	9	10	11	12
1 Brazil	--											
2 Congo	4.83	--										
3 Cuba	5.28	4.56	--									
4 Egypt	3.44	5.00	5.17	--								
5 France	4.72	4.00	4.11	4.98	--							
6 India	4.50	4.83	4.00	5.83	3.44	--						
7 Israel	3.83	3.33	3.61	4.67	4.00	4.11	--					
8 Japan	3.50	3.39	2.94	3.83	4.22	4.50	4.83	--				
9 China	2.39	4.00	5.50	4.39	3.67	4.11	3.00	4.17	--			
10 Russia	3.06	3.39	5.44	4.39	5.06	4.50	4.17	4.61	5.72	--		
11 U.S.	5.39	2.39	3.17	3.33	5.94	4.28	5.94	6.06	2.56	5.00	--	
12 Yugoslavia	3.17	3.50	5.11	4.28	4.72	4.00	4.44	4.28	5.06	6.67	3.56	--

[a]Larger numbers indicate greater similarity.

Table 5

Simulated Profiles for 21 Objects and 3 Variables

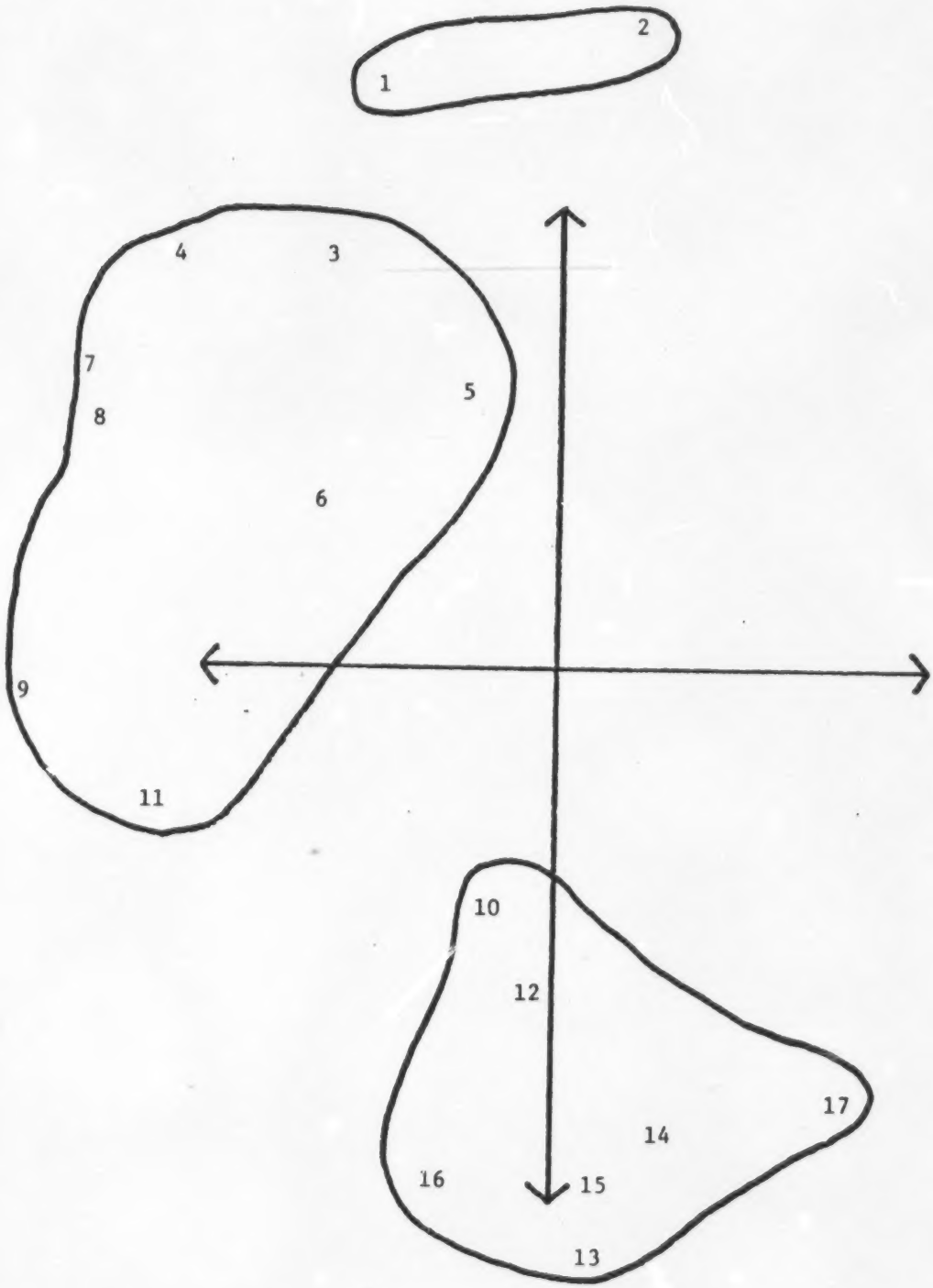
Subgroup	Object (i)	Y_{i1}	Y_{i2}	Y_{i3}
G_1	1	20	31	26
	2	17	29	25
	3	16	30	27
	4	18	31	28
	5	15	28	25
G_2	6	13	24	20
	7	13	27	21
	8	15	26	21
	9	15	27	18
G_3	10	23	32	29
	11	22	31	30
	12	22	36	32
	13	23	36	30
	14	22	33	33
	15	21	35	31
	16	25	35	33
	17	24	34	30
G_4	18	16	28	24
	19	18	28	26
	20	18	29	25
	21	20	31	25

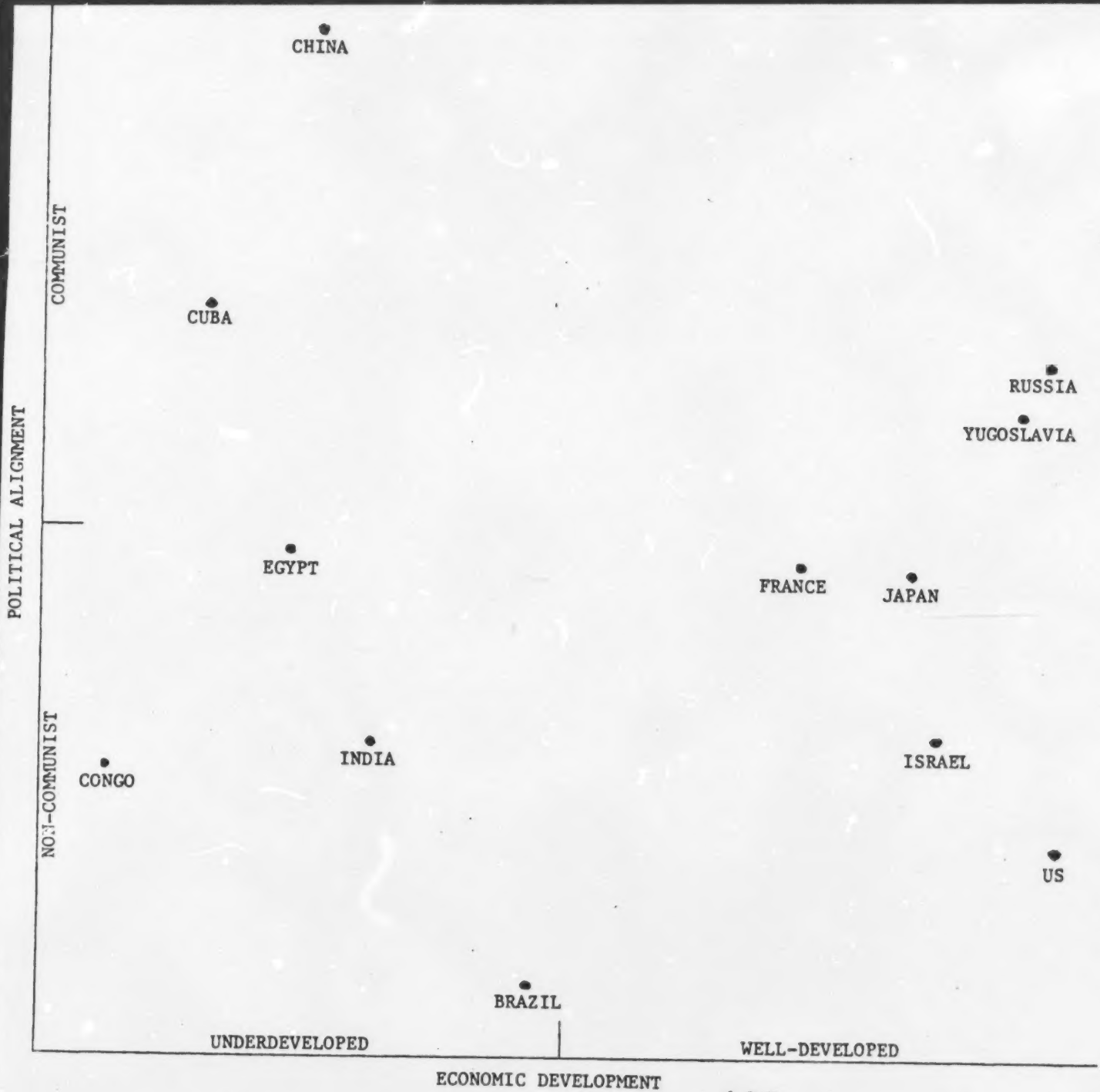
Figure Captions

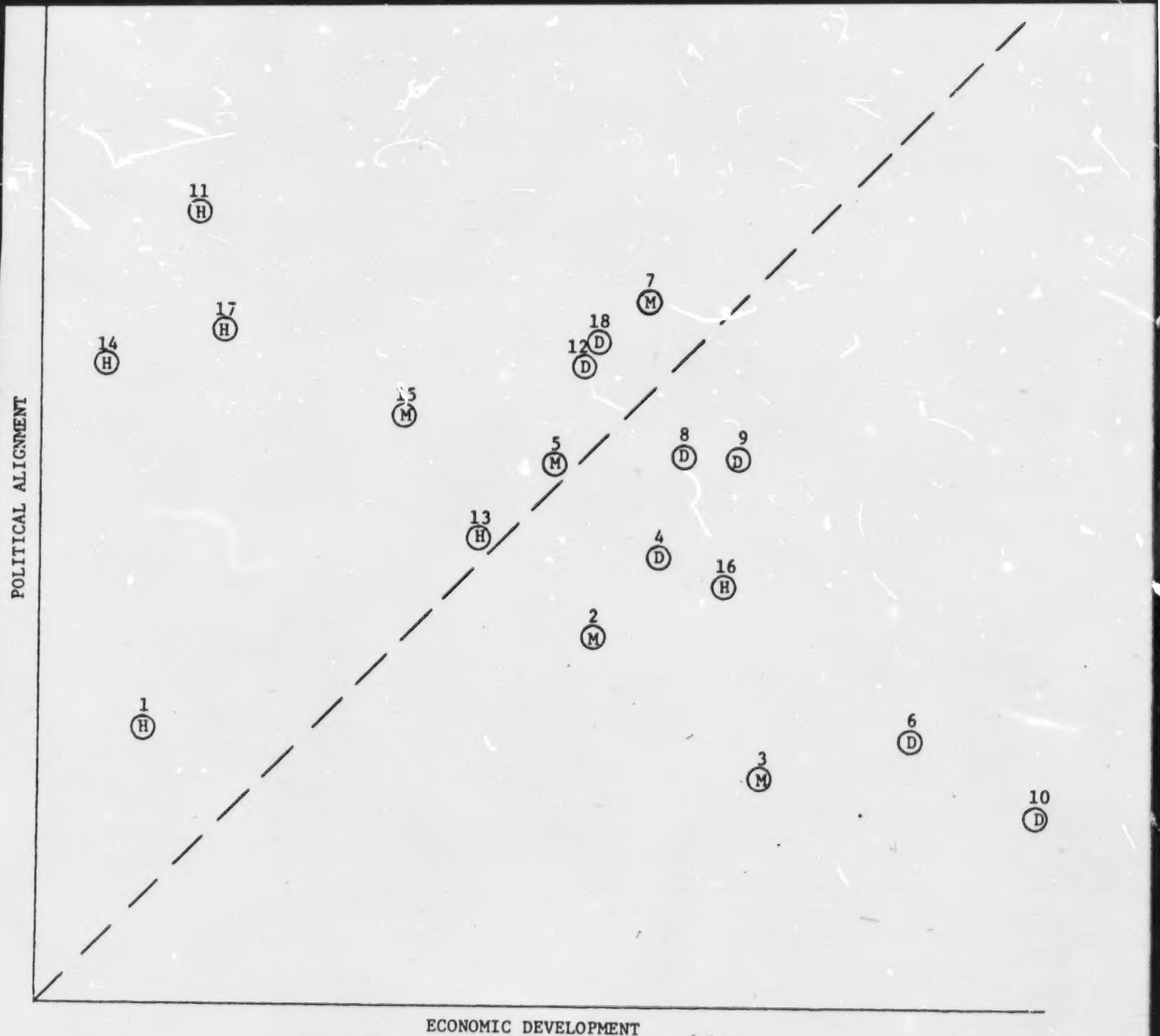
Figure 1. Two-dimensional scaling of the Glushko patterns.

Figure 2. Two-dimensional configuration of 12 nations.

Figure 3. Two-dimensional configuration of 18 subjects.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17


CHINA
CUBA
RUSSIA
YUGOSLAVIA
EGYPT
FRANCE
JAPAN
INDIA
ISRAEL
CONGO
US
BRAZIL
POLITICAL ALIGNMENT
COMMUNIST
NON-COMMUNIST
UNDERDEVELOPED
WELL-DEVELOPED
ECONOMIC DEVELOPMENT


POLITICAL ALIGNMENT
ECONOMIC DEVELOPMENT
11 H
17 H
14 H
7 M
18 D
12 D
15 M
5 M
8 D
9 D
13 H
4 D
16 H
2 M
1 H
6 D
3 M
10 D

Reference Note

1. Carroll, J. P. & Chang, J. J. A general index of nonlinear correlation and its application to the problem of relating physical and psychological dimensions. Unpublished manuscript available from Bell Telephone Laboratories, Murray Hill, New Jersey (no date).

References

Althauser, R. P., Burdick, D. S., & Winsborough, H. H. The standardized contiguity ratio. Social Forces, 1966, 45, 237-245.

Campbell, D. T., Kruskal, W. H., & Wallace, W. P. Seating aggregation as an index of attitude. Sociometry, 1966, 29, 1-15.

Carroll, J. D., & Chang, J. J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika, 1970, 35, 283-319.

Cliff, A. D., & Ord, J. K. Spatial autocorrelation. London: Pion, 1973.

Garner, W. R. Uncertainty and structure as psychological concepts. New York: Wiley, 1962.

Geary, R. C. The contiguity ratio and statistical mapping. The Incorporated Statistician, 1954, 5, 115-141.

Glushko, R. J. Pattern goodness and redundancy revisited: Multidimensional scaling and hierarchical clustering analyses. Perception and Psychophysics, 1975, 17, 158-162.

Hubert, L. J. Nonparametric tests for patterns in geographic variation: Possible generalizations. Geographical Analysis, 1977, in press.

Hubert, L. J., & Baker, F. B. Analyzing distinctive features. Journal of Educational Statistics, 1977, 2, 79-98.

Hubert, L. J., & Levin, J. R. Evaluating object set partitions: Free-sort analysis and some generalizations. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 459-470. (a)

Hubert, L. J., & Levin, J. R. A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin, 1976, 83, 1072-1080. (b)

Hubert, L. J., & Schultz, J. V. Quadratic assignment as a general data analysis strategy. The British Journal of Mathematical and Statistical Psychology, 1976, 29, 190-241.

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

Kaiser, H. F. A second generation little jiffy. Psychometrika, 1970, 35, 401-415.

Kruskal, J. B. Multidimensional scaling: A numerical method. Psychometrika, 1964, 29, 1-27. (a)

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 1964, 29, 115-129. (b)

Mantel, N. The detection of disease clustering and a generalized regression approach. Cancer Research, 1967, 27, 209-220.

Mielke, P. W., Berry, K. J., & Johnson, E. S. Multi-response permutation procedures for a priori classifications. Communications in Statistics--Theory and Methods, 1976, 5, 1409-1424.

Royaltey, H. H., Astrachan, E., & Sokal, R. R. Tests for patterns in geographic variation. Geographical Analysis, 1975, 7, 369-395.

Schultz, J. V., & Hubert, L. J. A nonparametric test for the correspondence between two proximity matrices. Journal of Educational Statistics, 1976, 1, 59-67.

Winsborough, H. H., Quarantelli, E. L., & Yutzy, D. The similarity of connected observations. American Sociological Review, 1963, 28, 977-983.

Wish, M., Deutsch, M., & Biener, L. Differences in conceptual structures of nations: An exploratory study, Journal of Personality and Social Psychology, 1970, 16, 361-373.

Footnote

Equal authorship is implied. Lawrence Hubert and Michael Subkoviak were supported respectively by grants from the National Science Foundation (SOC 75-07860) and the National Institute of Education (NIE-G-76-0088).

Requests for reprints should be addressed to Lawrence Hubert, Department of Educational Psychology, The University of Wisconsin, Madison, Wisconsin, 53706.